VÁCLAV BELÁK

A STRUCTURAL APPROACH TO COMMUNITY-LEVEL SOCIAL INFLUENCE ANALYSIS

A STRUCTURAL APPROACH TO COMMUNITY-LEVEL SOCIAL INFLUENCE ANALYSIS

VÁCLAV BELÁK

Ph.D. Thesis

Supervisor: Dr. Conor Hayes
Co-supervisor: Dr. Marcel Karnstedt

External Examiner
Prof. Dr. Mike Thelwall

Internal Examiner
Prof. Dr. Stefan Decker

Insight Centre for Data Analytics
College of Engineering and Informatics
National University of Ireland, Galway

January 2014

ABSTRACT

Social communities shape the way people interact. E. g. members of online discussion communities frequently exchange information, experience, or knowledge about practically anything from software to bird watching. The rising availability of data from social communities has led to a surging research interest in their modelling and analysis. One of the main motivations behind the interest is the promise that models of community dynamics may help the stakeholders to make good use of the time or capital they invest to the communities.

A prominent problem in the study of communities has been to quantify and explain how their members *influence* each other. This has found many applications in various areas such as public health promotion programs or viral marketing. However, how a community, as entity, influence or is influenced by other communities, i. e. *cross-community influence*, has been less studied. We propose that the relationships a community, as a whole, maintains with other communities contribute to how it evolves in terms of growth, topic, and decline.

The *main problem* we address is the measurement, analysis, and explanation of influence relationships between various types of social communities. We address the problem by developing a computational model for *c*ross-*co*mmunity *in*fluence that we call COIN. Our model is flexible and caters for differences between data from various types of communities. The core of COIN is based on a purely network-based representation of social interactions. COIN can thus reveal and explain influence relations between communities for which no additional data like textual content is available due to e. g. legal reasons. However, we also devise an extended version that integrates and helps to interpret additional information about the interactions extracted from textual data.

The model is evaluated on three data-sets from leisure, business, and scientific communities. We present and explain a broad range of cross-community influence phenomena. We describe a rise of global authorities or communities that act as a hub, as well as dynamic patterns of influence between pairs of communities. Furthermore, we demonstrate how can be the cross-community influence exploited for efficient information diffusion. Last but not least, we use COIN to identify scientific communities that became increasingly isolated, self-referential, and shrinking, thus shedding more light onto the possible causes of a community's decline.

I give all the merit that may arise from this work to my parents:
Všechny zásluhy, které případně vzejdou z této práce, věnuji mým rodičům:
Ireně a Zbyňkovi.

ACKNOWLEDGMENTS

I have had the honour to meet and work with many remarkable people whom I am grateful for their support, wisdom, and energy. First of all, I thank to my adviser, Dr. Conor Hayes, who has taught me countless aspects of doing research, and write and think as a scientist. Likewise, I thank to my second adviser, Dr. Marcel Karnstedt, who has "showed me the ropes" and always offered help and advises. It has been their friendly and energetic guidance that helped me to overcome several of the hard moments of my research. Furthermore, I thank to Prof. Stefan Decker, and Dr. Adrian Mocan from SAP for their insightful suggestions. However, I would like to emphasise that any possible shortcomings of this thesis are fully my own responsibility.

I have been lucky to be surrounded and supported by a loving and fun family that has not let me to forget that there is life beyond research. I particularly thank to Weronika and Staś for their love, patience, and understanding.

Last but not least, I would like to thank to the community that has influenced my life like no other community: the people from the former Digital Enterprise Research Institute, who have made the past four challenging years incredibly interesting and fun for me. I particularly thank to (in random order): Jeff, Sam, Donn, Ioana, Benjamin, Erik, Jodi, and John for being the best colleagues I've ever had; Vít & Hanka for all the fun; Josef & Hanka for all the fish; Hugo for all of his admin magic; Hilda, Clare, Christiane, and Carmel for their kindness and help; Jakub, Pierre Ludwick, Myr, and Fabrizio for all the vibe; Andrejs for all the great rope work, and last but not least Ger and Andrew for maintaining the best infrastructure I've ever used.

Finally, I sincerely thank to all Irish and EU taxpayers who have supported my research. In particular, I have been supported by:

- the Science Foundation Ireland (SFI) under Grant No. 08/SRC/I1407 (Clique: Graph & Network Analysis Cluster) and SFI/12/RC/2289 (Insight)
- the European Union under Grant No. 257859 (ROBUST)

CONTENTS

ACRONYMS

AI Artificial Intelligence

CAICM Community-Aware Independent Cascade Model

CALTM Community-Aware Linear Threshold Model

CIF Conference Impact Factor

COIN Cross-Community Influence Analysis Framework

CBR Case-Based Reasoning

GI Group In-degree

GR Greedy targeting strategy

ICM Independent Cascade Model

IF Impact Focus

LTM Linear Threshold Model

RA Random targeting strategy

SYMBOLS

a	user activation fraction
c	community activation fraction
P	set of documents
k	number of communities
d	number of topics
m	number of the modes of a tensor
n	number of actors
$\mathbf{s}$	vector of community sizes
$\mathbf{W}$	weight matrix
$\mathbf{M}$	membership matrix
$\mathcal{M}$	membership tensor
$\mathbf{C}$	centrality matrix
$\mathcal{C}$	centrality tensor

$\mathbf{J}$	impact matrix
$\mathcal{J}$	impact tensor
$\mathbf{S}$	community sizes matrix
$\mathcal{S}$	community sizes tensor
τ	cross-posting waiting time
t	time
θ	threshold
N	set of neighbours
s	seed actors sample size
q	number of seed communities
R	overall number of replies between two actors
z	transmission probability
B	number of replies an actor contributed to a community
F	set of replies received by a user within a community
T	set of seed communities
L	set of seed actors
A	set of all the actors that have been activated
C	set of all the members of a community or a random variable representing communities
σ	mapping from the set of seed actors to all the actors that have been activated
h	number of repetitions
D	random variable representing documents
Z	random variable representing topics
p	probability distribution
V	set of actors
E	set of ties
G	network (graph) of actors connected by ties
φ	membership (characteristic) function of a set
ψ	actor similarity metric

DECLARATION

I declare that this thesis is composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification.

Galway, January 2014

Václav Belák

1 INTRODUCTION

Social communities represent a natural and important organisational frame of human interactions. For example, scientists engage with groups of their peers from the same discipline at conferences, or people exchange information and knowledge in online discussion or question-answering communities. In addition, the success of online communities is expected by Gartner to propagate into enterprise in the next three years [62]. The increasing availability of data about social communities and their members has led to a surging interest in their research and analysis [91, 92, 53]. One of the reasons is that analysis and modelling of social communities may help their stakeholders to better understand, monitor, or even design and manage [53] their communities. Therefore, the research of social communities promises to help the stakeholders to make good use of the time or capital they invest to the communities.

A common approach to study social communities and their members is to represent the individual interactions between the members, or *actors*, as a *social network*. The *structural approach* to analysis of human interactions proposes that a social network characterises *flow* of some material or non-material *resources* between actors or their groups [102, p. 4]. For example, frequent communication between people can be analysed as an information flow network.

A prominent problem in structural analysis is to find the actors who are *influential* in some sense [102, p. 169], [34, 92, 51, 3, 10], e. g. those who are in control of how information flow over the network. Actor-level influence analysis has found many important applications in diverse domains like public health [100], marketing [3], or innovation management [99]. However, how communities, as entities, influence and are influenced by each other, i. e. *cross-community influence*, has received little attention [102, p. 202].

The *main problem* we address in this thesis is the measurement, analysis, and explanation of cross-community influence in different types of dynamic social communities. By addressing this problem, we deliver a crucial insight into how communities are shaped by external relationships as much as they are shaped by the internal relationships among their individual members. We demonstrate that the relationships a community, as a cohesive whole, maintains with other communities significantly contributes to how it evolves in terms of growth, topic, and decline.

The influence relationships between communities are induced by the interactions between their individual members and, naturally, some members may be shared between two or more communities. While people may contribute to multiple communities, there are typically a few communities to which they are strongly affiliated and regularly contribute to. The *focal community* is the place in which an actor mostly participates and, by definition, to which she strongly belongs. Conversely, *alter communities* are places where the actor's participation is infrequent. Each community is thus composed of strongly committed *focal members* and less committed *alter members*. However, an actor may also have more than one focal community or, conversely, she may not belong strongly to any community altogether. By quantifying the distribution of actors' activity and interaction we generate insight into the pattern of influence and dependencies that any community may exhibit.

The underlying behaviour in many social communities is centred around information exchange between the individual members. Actors contribute information to communities and an actor may *respond* to another actor. For example, users of online discussion communities contribute messages that are often in response to one another. Likewise, scientists publish papers that usually cite other papers. The responses can be represented as a social network, where a link connects a responding actor to the responded actor. Therefore, the dynamics of communities and their interactions can be investigated by structural analysis of the social network that underpins the communities.

In many cases, an actor who stimulates many responses can be deemed as *important*, because she *influences* her peers towards high *activity*. For example, counts of incoming citations are often used as a basis for assessment of scholarly impact [71]. The influence or importance of an actor can be measured by a *centrality score* that characterises the actor's *position* in the network. For instance, an actor stimulating many responses will have many incoming links—a high *in-degree centrality* in the network. We put forward a *hypothesis* that the ability of focal members to stimulate activity in alter communities is a measure of a community's influence.

With this assumption, an *influential community* will be a community whose focal members have high centrality scores in a number of alter communities. Conversely, a *dependent community* is a community whose alter members have substantially greater centrality scores than focal members. In other words, the community is dependent on alter members to generate activity. This does not necessarily mean that it is weak—it may act as a sandbox for behaviour or topics that may be inappropriate in focal communities [53, p. 132] or it may play a particular social function. For example, a dependent community may serve as a *hub*, i. e. a common meeting place with broader focus, where focal members of communities with narrow focus meet. On

the contrary, high total influence of a community may indicate its authority in the system. This is similar to the node centrality measure HITS [52] that assigns hub and authority scores to a node depending on its position in the network.

We develop a structural approach to community influence that uncovers and quantifies the network of influence between dynamic social communities. The approach is formalised and implemented into a computational framework for *c*ross-*co*mmunity *in*fluence, COIN. In the remainder of the introduction, we first present the scope of this thesis and our main research questions. Afterwards, we present the structure of the thesis. Finally, we list the main contributions and the publications that resulted from our research.

1.1 SCOPE OF THE THESIS AND RESEARCH QUESTIONS

We assume communities whose members engage in regular exchange of information by responding to each other. We apply the framework on data-sets of three different types of communities. The first data-set spans 10 years of data of the largest online discussion system in Ireland, Boards.ie [17]. The second data-set represents 8 years of data of online technical support and question answering fora from SAP Community Network (SAP SCN) [88]. Finally, in addition to the two online systems, we analyse 19 years of citation data representing researchers and communities in computer science.

INFLUENCE AND IMPACT Social influence is generally understood in many ways as an ability of one actor to change opinions, behaviour, or emotions of other actors. The notion of influence that we adopt is specifically tied to the mutual interactions of actors in the form of responses. An influential community (or actor) is therefore stimulating other communities (actors). Hence, our notion of influence is *activity-based*. Throughout the thesis, we use the word *impact* as a *quantification* of the phenomenon of influence. The *influence* as a *qualitative* characteristic of a community is therefore indicated by a substantial and persistent impact of the community.

RESEARCH QUESTIONS We conducted four studies, described in Chapters 4–7, examining our hypothesis by extensive structure-based exploratory and qualitative analysis supplemented by automated content processing, and simulation experiments in which we address namely the following *research questions*:

Q1 How can we identify communities persistently influencing a particular community? How does the influence change over time?

Q2 Are there highly influential or dependent communities in the system? Do they coincide with authoritative or hub communities?

Q3 How can we find communities that are highly influential or influenced with respect to a particular topic?

Q4 Can we exploit the cross-community influence for efficient information or influence diffusion over the network?

1.2 STRUCTURE OF THE THESIS

In the following, we describe the structure of the thesis and we briefly introduce the topic of each chapter along with the research questions it primarily addresses. In addition, Figure 1 illustrates the possible ways that we recommend to proceed through the thesis.

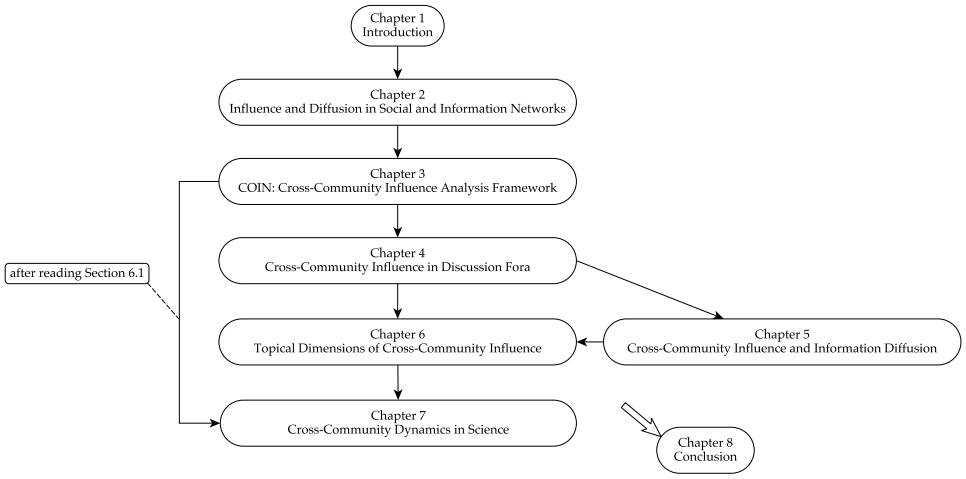


Figure 1: The possible ways that we recommend to proceed through the thesis.

CHAPTER 2 In the next chapter, we define our terminology; survey and critically review the related literature; and identify several limitations of the state-of-the-art.

CHAPTER 3 We address the fundamental limitations in Chapter 3. We motivate and present our hypothesis of structure-based cross-community influence and develop the hypothesis into a set of measures that constitute the *purely structural* core of COIN.

CHAPTER 4 We first evaluate COIN on data from online fora Boards.ie and SAP SCN (questions Q1 and Q2). We observed, for example, highly dependent communities acting as hubs, or, conversely, highly influential communities, that were often relatively small and private, thus revealing specific grouping behaviour of elite actors.

CHAPTER 5 We further apply COIN in the context of information diffusion. Previous studies typically aim to maximise the spread of information over a social network by engaging with a small set of initially stimulated *seed actors*. However, in many cases the information is communicated to the community as a whole. Therefore, the main problem we address in Chapter 5 is to maximise information diffusion starting from a small set of *seed communities* (Q4).

CHAPTER 6 In Chapter 6, we extend the purely structural framework from Chapter 3 in order to investigate the topics that may underpin the cross-community influence relations (Q3). We show that the extended COIN offers better interpretability and that it amplified the signal that we were able to extract from the data.

CHAPTER 7 The fourth and last of our studies investigates cross-community relations in communities of researchers in computer science (Q1 and Q2). We demonstrate, for instance, how COIN can enable identification and explanation of dynamics of scholarly communities that have become increasingly self-referential, isolated, and shrinking in size.

CHAPTER 8 Finally, in the last chapter we conclude the thesis, discuss the limitations of COIN, and we offer several perspectives on future research topics that have emerged out of our research.

APPENDIX A Appendix A details the preparation of the citation data that we analysed in Chapter 7.

ONLINE SUPPLEMENTARY MATERIAL All the software and data, along with additional outputs that we omit from the thesis for space or technical reasons are available online [12].

1.3 CONTRIBUTIONS OF THE THESIS

The main contribution of this thesis is that we provide a solution to the problem of measurement, analysis, and explanation of influence between a broad

range of social communities. We motivate, formally develop, and empirically evaluate a computational model for cross-community influence that we call COIN. By using COIN, we were able to reveal and systematically describe a broad range of cross-community phenomena in three different types of social communities: two online and one offline. Our findings demonstrate that COIN enables a rich set of insights into how communities evolve and influence each other in terms of their topics, activity, growth, and decline. For those reasons, COIN was adopted by SAP corporation in their product PULSAR (Pulse Check Application for Online Communities) [70].

1.3.1 *List of Main Contributions*

Specific contributions include:

- A *computational model* for quantification of dynamic cross-community influence. The core model is based purely on structural features and thus it generates many insights into cross-community dynamics even if no additional data like textual content is available due to e. g. legal or technical reasons (Chapter 3). We further devise an extended model that integrates additional information extracted from textual data (Chapter 6).
- An *empirical analysis* of the three types of communities that has generated insights into the emergence of highly influential or influenced communities. We believe that our analysis is an important step towards systematic understanding of cross-community influence phenomena in general (Chapters 4, 6, and 7).
- The first approach to the problem of *information diffusion maximisation* by selecting *seed communities*. We prove its NP-hardness and propose a greedy hill-climbing and a COIN-based heuristics for its solution (Chapter 5).
- We investigate cross-community dynamics of computer science communities and identify highly influential or highly self-referential communities (Chapter 7). Therefore, we show that apart from studying online discussion communities, COIN is a useful analytical tool in the fields like bibliometrics [97] or scientometrics [59].

1.3.2 *List of Publications Related to the Thesis*

The research presented in this thesis has been partially published in the following publications.

Research papers:

- Belák Václav, Lam Samantha, Hayes Conor. Cross-Community Influence in Discussion Fora. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'12)*. AAAI. 2012.
- Belák Václav, Lam Samantha, Hayes Conor. Towards Maximising Cross-Community Information Diffusion. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'12)*. IEEE/ACM, 2012.
- Belák Václav, Karnstedt Marcel, Hayes Conor. Life-Cycles and Mutual Effects of Scientific Communities. In *Procedia—Social and Behavioral Sciences*. ISSN 1877-0428. 22:37–48, 2011.
- Belák Václav, Karnstedt Marcel, Hayes Conor. Life-Cycles and Mutual Effects of Scientific Communities. Presented at the *Conference on Applications of Social Network Analysis (ASNA'10)*. Zurich, Switzerland, 2012.
- Belák Václav, Lam Samantha, Hayes Conor. Targeting Online Communities to Maximise Information Diffusion. In *Proceedings of the International Conference companion on World Wide Web (WWW'12)*. ACM. 2012.

Posters:

- Belák Václav, Lam Samantha, Hayes Conor. Cross-Community Influence Analysis and Maximisation. Presented at *The International School and Conference on Network Science (NetSci'12)*. Evanston, IL, USA, 2012.
- Belák Václav, Lam Samantha, Hayes Conor. Cross-Community Influence Analysis and Maximisation. Presented at *The Annual Research Day of NUI Galway*. Galway, Ireland, 2013. **Best Poster Award** in the category IT/Mathematics.
- Belák Václav, Karnstedt Marcel, Hayes Conor. Life-Cycles and Mutual Effects of Scientific Communities. Presented at *The Royal Society Web Science Meeting*. Kavli Royal Society International Centre, UK, 2011.

2

INFLUENCE AND DIFFUSION IN SOCIAL AND INFORMATION NETWORKS

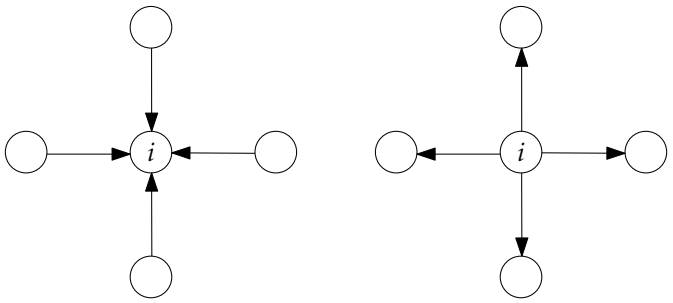
Networks in general are a very powerful framework for representing many different complex systems and not only social interactions [76]. Many of the advances in the analysis of networks thus quickly spread beyond the particular domain they were developed in. For example, some algorithms that were proposed for measurement of authoritativeness of pages on the Web were inspired by methods for prestige measurement in social network analysis or by indicators of scholarly impact developed within bibliometrics. These algorithms then have been subsequently adopted by the researchers and practitioners in social network analysis or bibliometrics. The ascent of the Web increased the interest in the structural analysis also for another reason. The networks that had to be previously laboriously obtained, observed, or reconstructed, became often *directly observable* on the Web. The research of impact and influence in networks thus represent an exemplary collaboration feedback between fields like sociology, bibliometrics, statistical physics, and computer science to name but a few.

In this chapter, we review the contributions in study of influence and information diffusion that are fundamental for our research of cross-community influence. First, we establish the core terminology in Section 2.1. After that, we proceed with reviewing the essential concepts of social network analysis in Section 2.2. A research of influence between individuals or their groups has a long tradition in analysis of citation networks and in bibliometrics as we present in Section 2.3. In Section 2.4 we show that bibliometrics and social network analysis influenced strongly how we understand and use the Web today. The rise of the Web, however, had an enormous effect on the way people interact in all aspects of modern life. This has lead to non-traditional or cyber communities organised around shared interests, goals, or affiliations. Therefore, we review the core contributions in research of influence and information diffusion on the Web, with a particular focus on online discussion communities in Section 2.5. Finally, in Section 2.6 we summarise the contributions in research of cross-community influence and their limitations.

2.1 ELEMENTARY TERMINOLOGY

Before we proceed with our survey, it is useful to establish the core terminology. The central concept of the structural approach is a *network* of individual people or their groups linked by means of their mutual relationships. The relationships represent flow of some material or non-material resources [102, p. 4], e. g. information in a form of regular exchange of messages. If the individuals in the network represent people, we refer to them as *actors* and to the links between them we refer to as *ties* [102, p. 17–18]. We denote a tie from actor i to actor j as an ordered pair (i, j). Finally, we can represent the entire network of a set V of n actors connected by a set of ties $E \subseteq V \times V$ as a *graph* $G = (V, E)$ [102, p. 67], [66]. Please note that outside the context of social network analysis, the set V is commonly referred to as *nodes* or *vertices*, and the set E is often called *edges* or *links*.

The ties represent a *relation* between the actors. If the relation is symmetric, e. g. friendship, then the tie is called *undirected*. Otherwise, if the relation is not generally symmetric, e. g. if (i, j) represents that actor i sent a message to j, we say that the tie is *directed*. A set of actors N_i adjacent to actor i is frequently referred to as *neighbours*. If the ties are directed, we can differentiate two classes of neighbours as illustrated in Figure 2. *In-neighbours* N_i^{in} of actor i is the set of actors with a tie incoming into i, formally $N_i^{in} = \{j | (j, i) \in E\}$. Conversely, the set of actors to whom actor i is connected by an outgoing tie is called *out-neighbours* $N_i^{out} = \{j | (i, j) \in E\}$.



(a) in-neighbours (b) out-neighbours

Figure 2: The neighbours of actor i.

There are several types of networks (graphs) that frequently appear in the context of structural analysis. First, depending on the type of the relation that induces the ties, the network may be *directed* or *undirected*. Second, a tie may have an additional numerical attribute, a *weight* or *strength* of the tie, that characterises quantitatively the relation between the connected actors.

For example, it may represent the number of times the two actors connected by a tie have exchanged a message. In the cases like that, we say that the network is *weighted*. Both weighted and unweighted networks can be conveniently represented as an $n \times n$ *weight matrix* $\mathbf{W}$. The existence of a tie (i, j) is reflected by setting $\mathbf{W}_{ij}$ either to the value of the weight of the tie, or to $\mathbf{W}_{ij} = 1$ if the tie is not weighted. If there is no tie from i to j, then $\mathbf{W}_{ij} = 0$. Finally, depending on whether the structure of the network, i. e. its actors and ties, changes over time or not, we talk about *dynamic* or *static* networks.

We already mentioned that people often group together in communities centred around a shared interest, goal, or affiliation. The previous research suggests that communities are formed by actors who are in some sense similar to each other (e. g. based on their shared interests, goals, or affiliations) [91, p. 153]. Furthermore, the previous research suggests that communities facilitate formation of ties [33, 106] between the actors and therefore provide opportunities for influence and information flow [102, p. 297]. For example, a set of conference attendees form the conference's community, because conferences are usually focused on some specific domain and one of the main purposes of scientific conferences is indeed to provide platform for networking and communication. Finally, the frequent interactions of the community's members and their contribution to the community often leads to their long-term commitment to the community [53, p. 77]. The findings in social psychology suggest that the commitment fundamentally contributes to the sustainability of the community because the committed members generate more activity, help the newcomers, and care about the community even if it faces some hardship [53, p. 77]. Hence we adopt the following definition of a community:

Definition 1 *We define* community *u as any non-empty subset C_u of the set of actors V,*

- *who are in some sense similar to each other as measured by a* similarity function *$\psi : V \times V \to \mathbb{R}^+$;*
- *and who have some sense of belonging to the community, represented as the characteristic or* membership *function $\varphi_u : V \to [0, 1]$ of the set C_u, i. e. $C_u = \{i | \varphi_u(i) > 0\}$.*

The membership function φ_u maps any actor to a number quantifying the membership of the actor within the community. Depending on the range of the membership function, we differentiate *crisp* and *fuzzy* communities [42]:

Definition 2 *We say that community u is* crisp *if:*

- *it meets the requirements from Definition 1;*
- *the* membership *function φ_u of the set C_u ranges only within $\{0, 1\}$.*

Therefore, actor i is either fully a member of crisp community u and then $\varphi_u(i) = 1$, or she is not a member and then $\varphi_u(i) = 0$. In other words, the set C_u is *crisp*. Alternatively, if the degree of belonging of the members of community u varies within $[0, 1]$, the community is formed by a *fuzzy* set [110]:

Definition 3 *We say that community u is* fuzzy *if:*

- *it meets the requirements from Definition 1;*
- *the* membership *function φ_u of the set C_u ranges fully within $[0, 1]$.*

Clearly, the crisp representation is a special case of fuzzy communities where an actor either fully belongs to the community or not. The fuzzy representation of communities is more realistic whenever the memberships of an actor in two communities differ. For example, if actor i participates in two online discussion communities u and v, but overall she prefers to contribute to community u than to v, we can represent the difference in her preferences by a difference in her memberships, i. e. $\varphi_u(i) > \varphi_v(j)$.

Given the graph $G = (V, E)$ and the set of members of community u, we can *induce* a subgraph G_u by keeping only the actors and ties from the community. Formally,

Definition 4 *A* subgraph G_u *induced by community u from graph $G = (V, E)$ is defined as $G_u = (C_u \cap V, \{(i, j) \subseteq E | i, j \in C_u\})$.*

This is illustrated in Figure 3 where the subgraph inducted by community u is within the circle. Most of the methods for identification of *latent* or *implicit* communities in networks define a community as a densely connected subgraph. Therefore, these methods typically learn the membership function from the structure of the network [33, 106].

Alternatively, a community may be defined *explicitly* as a set of actors who deliberately joined some group, e. g. online discussion forum; or whose affiliation is known, e. g. university department; or who participate regularly in some *events*, e. g. scientific conferences. For example, the actors who

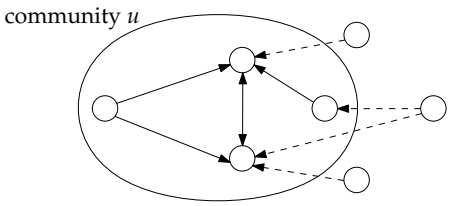
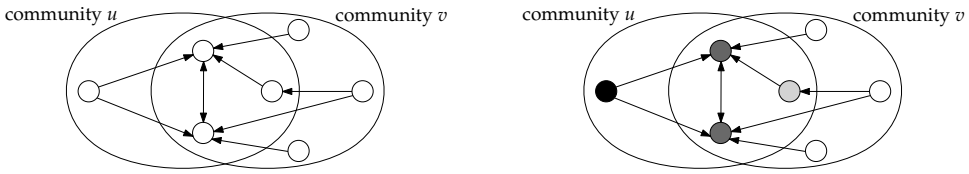


Figure 3: Community u as a subgraph induced by the set of its members C_i. The solid ties depict the subgraph. The dotted ties represent the rest of the network *outside* of the subgraph.

visit regularly an online discussion forum may be considered the forum's community. This is because an online discussion forum is usually centred around some topic that is discussed in the forum and thus the members of the forum's community are similar in terms of the shared interest in the topic. Furthermore, the frequent participation within the forum increases the likelihood that the actor replies—forms a *reply tie*—to some of the other participants. Finally, an actor that contributes a large amount of her time to the community is arguably committed to the community. Therefore, the membership function may be estimated by using the distribution of activity or time that the member contributed to the community as a proxy [68, 80].

Two or more communities can share some of their members—they *overlap* [42, 2, 106, 79, 107]. If the overlapping communities are crisp (Definition 2), then the shared actors belong to both communities equally [42] and we say that the overlap is *crisp*. Otherwise, if the communities are fuzzy (Definition 3), the actors at their intersection may belong to one community more than to the other. In such cases, we say that the overlap is *fuzzy*. As an example, consider two communities u and v as illustrated in Figure 4. We depict the conceptual difference between crisp (Figure 4a) and fuzzy (Figure 4b) overlaps by representing the membership using different shades of grey. Since the membership of actors in crisp overlapping communities from Figure 4a is equal to 1, all the actors have the same colour. Conversely, in the case of fuzzy overlap depicted in Figure 4b, the more grey (black) is the actor, the more she belongs to community u, and the lighter (whiter) she is, the more the actor belongs to community v.

In addition to actor overlap, in many cases a careful attention has to be paid to whether two or more communities can *share ties* between their actors or not. Let us now for illustration assume that Figure 4b represents two communities u and v of users of social networking site Google+. Each community corresponds to a social circle explicitly defined by some user. In this case, the ties between the overlapping actors are shared between the commu-



(a) crisp overlap (b) fuzzy overlap

Figure 4: Crisp and fuzzy overlaps between two communities u and v.

nities, because a friendship tie between two users in Google+ is formed by a direct "friending" action of the users and without additional information it cannot be exclusively assigned to either of the two communities. Therefore, the structure of community u can be represented as an induced subgraph depicted in Figure 3. In other cases, however, the ties may be specific to a community. For example, in many online discussion communities an actor can reply only to messages that were contributed to the same community. Hence the reply ties are not shared but they belong to the community where the reply was contributed. In order to reflect this, we define a stricter notion of the induced subgraph:

Definition 5 *A* confined subgraph *induced by community u is a subgraph induced by the community (see Definition 4) that contains only the ties that were formed within the community.*

SUMMARY We have defined several key notions of structural analysis. The structural approach represents a social system as a *network* or *graph* of *actors* linked by their *ties*. The ties can have a *weight* that quantifies the relationship between the connected actors. Depending on whether the network changes over time or not, it is either *dynamic* or *static*. A set of actors who are in some sense similar and whose similarity increases the likelihood of their interaction is called a *community*. When the community is defined in terms of its structure as a densely connected subgraph, we refer to it as an *implicit community*. Otherwise, the composition of an *explicit community* is known directly from actors' affiliations or participation in some events. If two communities share some members, we say that they *overlap*. We distinguish two types of overlaps: crisp and fuzzy. *Crisp overlap* assumes that an actor belongs to all her communities with constant membership. The more general notion of *fuzzy overlap* allows the membership of an actor to vary. The structure of a community is often represented as a network of the community members. If the communities are overlapping and the ties can be shared between the communities, e. g. friendship ties, the structure of a community

can be represented as a *subgraph* induced by the community. However, if a tie always belongs to a particular community, e. g. a *reply tie* in discussion communities, the structure of a community is represented by a *confined subgraph*. As we discuss in the rest of the chapter, the modelling assumptions behind the structure of the communities—whether they overlap and how—fundamentally affect the applicability of the methods that have been proposed previously for measurement of cross-community influence.

2.2 INFLUENCE AND SOCIAL NETWORK ANALYSIS

As we already indicated in Chapter 1, one of the core problems in structural analysis is to find the actors that are in some sense highly influential, important, or prominent [102, p. 169]. The problem of social influence quantification has been studied very intensively and many network-based measures have been proposed [34, 92, 43], [102, p. 169]. In this section, we focus only on the literature that is either directly related to cross-community influence, or that represents the theoretical foundations for our work. First, we discuss the measurement of actor influence and the related notion of *actor centrality*. After that, we review the literature on group-level centrality and cross-community influence.

2.2.1 *Actor Centrality and Influence*

Since the ties represent *flow* of either material or non-material *resources*, an actor, whose *position* in the network enables her to control, influence, or in any other highly benefit from the flow can be considered as influential or important. We call a measure that characterises the *influence* or prestige of a position of an actor in the network the actor's *centrality* [102, p. 172].[1] A simple yet powerful measure of an actor's centrality is the number of her neighbours [51]. An actor with many neighbours is highly visibly to her peers and therefore has many opportunities to influence them. For example, a user of Google+ social networking site with many friends (neighbours) has a broad audience for the content she shares. Likewise, an actor that is a subject of many *incoming* ties is often influential. For example, in a network of citations between scientists, an actor that is highly cited is likely highly re-

1 Please note that some authors [102] use the term "centrality" only for undirected networks and "prestige" for directed networks. For the sake of simplicity, we use the term "centrality" for both types of networks.

garded by her peers. The *in-degree centrality* of actor i is therefore defined [76, p. 169], [102, p. 202] as the total number (or strength) of her incoming ties:

$$\text{in-degree}(i) = \sum_{j \in N_i^{in}} \mathbf{W}_{ji}, \tag{1}$$

where $\mathbf{W}$ is the weight matrix and N_i^{in} is the set of in-neighbours as defined in Section 2.1 and illustrated in Figure 2a.

Naturally, the type of relation between the actors affects how the centrality can be interpreted [102, p. 174–175]. If the tie (i, j) represents for example the fact that actor i advises (is an adviser of) actor j, then the actor that has many *outgoing* ties is likely to be influential. Analogously to the previous measure, the out-degree centrality of actor i is defined as the total number (or strength) of her outgoing ties [76, p. 169]:

$$\text{out-degree}(i) = \sum_{j \in N_i^{out}} \mathbf{W}_{ij}, \tag{2}$$

where $\mathbf{W}$ is again the weight matrix and N_i^{out} is the set of out-neighbours as defined in Section 2.1 and illustrated in Figure 2b.

The influence of an actor may be higher if she maintains ties with highly influential actors who are themselves highly influential and so forth. In fact, some of the fundamental algorithms for quantification of authoritativeness of Web pages or measurement of scholarly impact are based on such a recursive definition of influence as we describe in Sections 2.3 and 2.4. Such measures may also be used to quantify an actor's centrality too.

2.2.2 *Group Centrality and Influence*

Everett and Borgatti [31] generalised several centrality measures to groups of actors, moving the scope of their analysis to the actor-community level. For instance, they defined group degree centrality "as the number of non-group nodes that are connected to group members". Following their definition, we may quantify the *group in-degree* of community u as the number (or total strength) of incoming ties from outside of the community to the members of u:

$$\text{gi}(u) = \sum_{i:(i,j) \in E \wedge i \notin C_u \wedge j \in C_u} \mathbf{W}_{ij}, \tag{3}$$

where C_u is the set of actors from community u. For example, consider a group in-degree of community u in an unweighted network as illustrated in Figure 5. The group in-degree of community u is 4 because that is the

number of ties from outside of the community to the community members. Another approach is to aggregate actor-level measures like actor in-degree by averaging over all the members of the community [102, p. 170]. In either case, the resulting measure quantifies the position of the group members as a whole relative to the non-members. In other words it captures a relation between a group of actors and the rest of the network but *not* between two or more groups. Therefore, these methods do not represent relations between communities as entities.

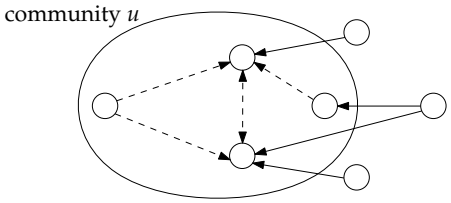


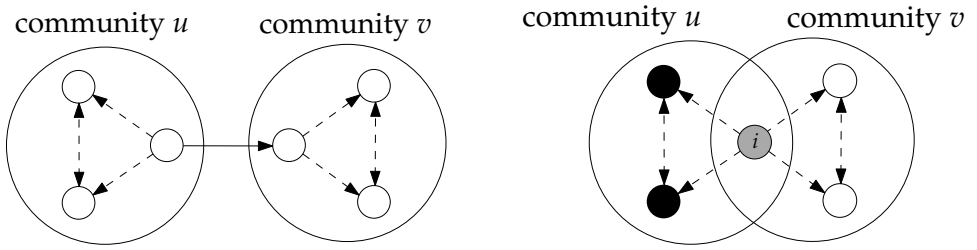
Figure 5: Group in-degree of community u as the number or total strength of the ties from outside of the community to the members of u. The ties from the outside are depicted as solid links whereas the internal ties of the community are illustrated as dashed lines.

In contrast with that, Friedkin [34, p. 169] proposed to quantify influence between academic communities as an average of inter-community influence ties between the individual actors. According to his model, the network of influence between the individual actors is derived first. Figure 6a illustrates an example network. A directed link represents that the actor at the source of the link influences the actor at the end denoted by an arrow. In the second step, the actors are clustered into *non-overlapping* communities as depicted in Figure 6a. The cross-community influence of community u on community v is then obtained as a mean of the influence ties, depicted as the solid line, from members of u to actors from community v.

The method proposed by Friedkin has several limitations. First, the author evaluated the method on communities of university employees only and therefore its applicability on other systems remains an open question. Second, the proposed method was tailored for non-overlapping communities. Therefore, it is not suitable for communities that overlap, because it is not clear how to aggregate influence ties among the overlapping actors. We illustrate this by the example of two fuzzy overlapping communities u and v in Figure 6b. We see that actor i at the intersection of the two communities influences the rest of their members. We further see that there are no other influences from a member of one community to a member of the other community. Since there are no influence ties from members exclusively be-

longing to u to the members exclusively belonging to v, the cross-community influence cannot be measured by the Friedkin's method.

However, if actor i belongs more to either of the two communities, say u, it may indicate that community u influences v. This is because actor i influences the other two members of community v, but she belongs more to community u. The structural approach that we propose in this thesis address these shortcomings by integrating information about distributions of the actors' memberships and influence.



(a) influence between non-overlapping communities

(b) influence between fuzzy overlapping communities

Figure 6: An illustration of the applicability of the cross-community influence measure proposed by Friedkin (see the main text).

Both group in-degree and the approach proposed by Friedkin do not allow to represent a *context* in which a cross-community influence may occur. Imagine, for example, a network of email communication between employees and their teams in some software development company. A team of database specialists are likely to frequently participate in communication about databases, but it is less likely to communicate about marketing. Therefore, the influence of a team may vary depending on a particular topic. Clearly, this would be revealed only by taking into account the additional information about the context like the textual content of the messages. Indeed, Spiliopoulou [91] in her recent survey identified integration of the content dimension as a one of the important motifs in research of social networks.

2.2.3 *Summary*

To sum it up, the influence of an actor is often quantified by her centrality, i. e. position, in a social network. An actor is thus typically considered influential if her position enables her to control or make use of resources that flow over the network. A simple yet powerful measures of actor centrality are in-degree and out-degree. The existing measures of community-level

centrality or influence are either defined on the level between communities and individual actors (e. g. group in-degree) but not on the level between individual communities; or they do not account for a possible overlap between the communities. Additionally, none of the methods for quantifying cross-community influence deal with additional data like content of messages that are contributed by the actors. Apart from social network analysis, research of influence and impact between individuals or their groups has a long tradition in bibliometrics (quantitative study of scientific and technical literature) [97], and scientometrics (quantitative study of science) [59].

2.3 CITATION NETWORKS AND BIBLIOMETRICS

Since the seminal work of Eugene Garfield on citation indexing [35], enormous effort has been devoted to research how scientists cite, i.e. refer to, each other in their articles, books, and other information artifacts. Even though the original intention of Garfield was to build an index for easier literature search, the citation index he and his colleagues have created, ISI (now part of Thomson Reuters), quickly became a fundamental element of many studies of citation networks. Probably the first [76, p. 68] was a paper by de Solla Price [27], who investigated a citation network of journals linked by the citations between the articles that they published. Since then, many different types of citation networks with actors representing individual researchers, their affiliations, or even whole countries have been studied [71]. Furthermore, the online publishing practises along with the advances in automated citation indexing [38] gave rise to large scale citation databases like CiteSeer, Google Scholar, and others [75]. The methods that have been proposed by bibliometricians for the analysis of citation data can be divided into either *evaluational methods* measuring impact, performance, or influence of the individual actors; or *relational methods* that aim to illuminate relationships between the actors [97].

2.3.1 *Evaluational Bibliometrics*

Mutual impact of individual scientists, their departments, institutions, or even the whole countries has been a subject of intensive research in evaluational bibliometrics [71, 97]. Most of the time the researchers analysed citation networks either between individual articles, or grouped to the level of journals, institutions, or countries. The meaning of a citation relation can be interpreted in multiple ways [71, p. 193]. In this regard, Martin and Irvine suggested to distinguish three characteristics of a publication: *quality*, *importance*, and *impact* [64, 63]:

QUALITY is a "property of the publication and the research described in it. It describes how well the research has been done, whether it is free from obvious 'error', how aesthetically pleasing the mathematical formulations are, how original the conclusions are, and so on."

IMPORTANCE is the publication's "*potential* influence on surrounding research activities—that is, the influence on the advance of scientific knowledge it would have if there were perfect communication in science ... However, there are 'imperfections' in the scientific communications system, the result of which is that the *importance* of a paper may not be identical with its *impact*."

IMPACT is the "*actual* influence on surrounding research activities at a given time. While this will depend partly on its importance, it may also be affected by such factors as the location of the author, and the prestige, language and availability of the publishing journal."

Citations as an Indicator of Impact

Clearly, the first two characteristics are hard to grasp by analysis of citation networks as they reflect different cultural, communication, and other biases [71, p. 204]. Martin and Irvine therefore argued, that it is the third characteristic, the impact, that is the most accurately indicated by citations. Moed suggested the term *citation impact* [71, p. 221] in order to emphasise the methodology according to which the impact is measured. However, the citation impact is only a *partial* indicator of the true impact, because apart from the impact of the paper, it is also influenced by the communication practises, the quality of data, the visibility of the authors, and so forth [63]. Therefore, the citation impact of actors from two different fields is generally incomparable due to the differences arising from the different publication practises or other biases. It is thus often recommended that in order to use citation impact for evaluational purposes meaningfully, only actors that are carefully matched according to their similarity, e. g. based on their discipline, should be compared by not just one, but according to multiple indicators [81, 63, 18, 97, 71].

Journal and Conference Impact Factors

The most widely known indicator is the *journal impact factor* [36] proposed by Garfield. In its original form, the journal's impact factor (JIF) in year t is defined as the average number of incoming citations a paper published in the journal in the preceding two years $[t-2, t-1]$ received in the year t. Since

its invention, there has been an intensive debate regarding its suitability, limitations, and extensions [71, p. 91], [86, 81, 65]. As in some fields, e. g. in computer science [101], the main medium of communication are not journals, but conferences, Martins et al. [65] proposed a *conference impact factor* (CIF). The definition of CIF is analogous to its journal counterpart, but the citations are aggregated on the level of conference proceedings instead of journals. One of the main limitations of JIF (or CIF) is that the differences in citation practises render the impact factors of two journals from distinct research fields generally incomparable [71, p. 95]. It is therefore suggested to either normalise the measures or compare only carefully selected samples of similar journals (conferences).

Citations as an Indicator of Influence

Another important limitation is that the citations from prestigious journals may indicate higher impact than the citations from the other journals. This was the fundamental observation of Pinsky and Narin [84], who proposed that a journal is influential if it is, recursively, highly cited by other influential journals. Geller [37] showed that this influence measure of a journal can be obtained as a stationary distribution using a following random walk on a network of journals linked by their mutual citations. Starting with an arbitrary journal, the walker selects a random outgoing citation and moves to that journal. Pinsky and Narin's notion of influence therefore corresponds to the time the hypothetical reader (walker) spends at journal. This has the advantage that even if the journal has relatively few citations, but if the citations are from highly cited journals, it is still deemed as influential. This method later inspired some of the fundamental information retrieval algorithms used for ranking Web pages, as we show in more detail in Section 2.4.

Citations as an Indicator of Information Flow

Another way how to estimate the strength of the impact the citation indicates is to directly measure how much information "flows" from the cited to the citing paper. Dietz, Bickel, and Sheffer [28] were able to estimate such flow of information between papers by calibrating a topic-relational graphical model that leverages both citation and language information. Other types of citations that do not express flow of information (e. g. "perfunctory citations" [61]) may reasonably be considered as a noise. This of course assumes that the proportion of citations that represent "information flow" to perfunctory citations is constant across the units of the analysis (e.g. researchers or their departments). While it seems plausible for a group of researchers in the same field, it is likely false for science in its entirety. Therefore, the citations

may be used as an indicator of information flow, but only for a carefully selected sample of publications that feature similar citation practises.

Citation Impact and Research Communities

An impact of one community of researchers on another can therefore be seen through the optics of information flow. Apart from the flow between the papers, information may also flow by collaboration between scientists. Montolio, Domingues-Sal, and Larriba-Pey [72] argued that the ability of a research community to attract and engage with new researchers is a good indicator of its performance. If for nothing else, at least in order to maintain its author base over the time. Even though the new members may have low membership in the community, they provide access to ideas and resources from other communities. Likewise, a group of researchers is more likely to have a higher impact if it is familiar with research outputs *beyond* its own boundaries. In this context Goldstone and Leydesdorff [40] talk about import and export of a community, i.e. the flow of ideas and knowledge to and from the community. An extreme case are the multi-university research teams that were reported to "produce the highest-impact papers if they include a top-tier university" [49]. In short, it seems that in addition to the performance of the individual members of a research community the impact of the community is also influenced by the position of the community as a whole in the network and by the relations that the community maintains with other communities.

2.3.2 *Relational Bibliometrics*

Apart from its usefulness for quantification of scholarly impact, the analysis of bibliographical data often illuminates relationships between the actors. The methods of relational bibliometrics therefore helps us to understand the general processes that drive the dynamics of science in general and how the impact of individual actors and communities come about. For example, *implicit* communities—or "hidden colleges"—can be identified by analysis of networks of researchers linked on the basis of their co-citation in reference lists [39]. As we noted in the beginning of this section, Garfield originally intended to use citations for indexing purposes. That is, a paper's references can be viewed as subject terms of the paper [71, p. 198–199].[2] Therefore, if two papers are frequently co-cited, they are likely related to similar subjects. Likewise, if two authors are frequently co-cited, they are supposedly part of the same or similar field. Figure 7 illustrates the way a co-citation link is de-

2 The invention of co-citation analysis is actually sometimes attributed to Garfield himself [97].

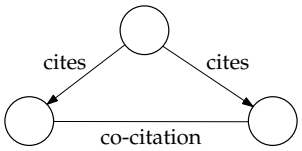


Figure 7: An example of how a co-citation link is derived from the citations between papers or their authors depicted as nodes.

rived from the citations. Unlike citations, co-citations are undirected because they represent similarity. Let us now briefly introduce our earlier exploratory co-citation analysis that triggered many of the research questions we address in this thesis.

The main aim of the study fully presented in our earlier paper [13] was to track and analyse patterns of interactions between implicit communities of researchers from two related fields of computer science: information retrieval (IR) and semantic web (SW). We picked these two fields because one of their dominant aims is to enable scalable and accurate information and knowledge retrieval on the Web, but their histories, methodologies, and author bases differ to a great extent. Hence we hypothesised that the flow of individual actors along with the topics they are associated with may reveal specific events such as an emergence of an interdisciplinary community on the boundary between IR and SW. We indeed observed that new communities associated with both IR and SW regularly emerge. Moreover, some of the newly emerged interdisciplinary communities seemed to have an important position bridging the communities focused predominantly on either SW or IR. Figure 8 illustrates one bridging community we identified. The observation that communities seem to interact and that the degree of their importance in the network seem to differ led us to the hypothesis that communities may have an impact on one another. However, this hypothesis cannot be pursued by investigating undirected networks like co-citations, because a co-citation link is an indicator of similarity, but not of impact (see Figure 7). Citation networks are therefore more suitable for illuminating impact relationships between scholarly communities.

For instance, Goldstone and Leydesdorff [40] mapped the relations and mutual influence between cognitive science and related fields such as psychology or artificial intelligence by investigating citation patterns between journals. A common approach adopted also by Goldstone and Leydesdorff is to assume that journals or their sets are proxy for the fields [15, 108, 80]. Therefore, the set of authors of a particular field defines *explicitly* its community. As we already noted earlier in this section, the main communication medium in computer science are conferences [101] and not journals. Conference proceedings thus represent an essential resource for any bibliometric

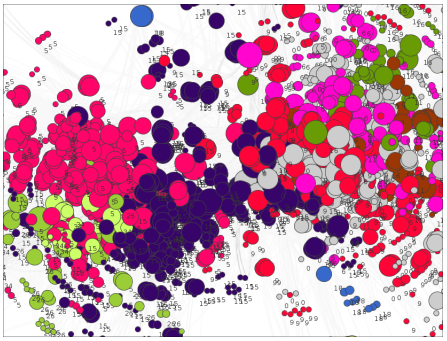


Figure 8: Communities of researchers from SW and IR. Nodes denote individual researchers and colours denote their community affiliations. The size of a node denotes the actor's betweenness [102, p. 189]. Please note the bridging community in the centre of the picture (violet). The communities on the left are predominantly centred around SW topics, whereas the topics of the communities on the right are primarily related to IR. Please see the paper [13] for more details.

analysis of computer science. It is probably the lack of suitable data that is the reason why, to the best of our knowledge, little work has been done in mapping the mutual influence between *explicit* computer science communities.

2.3.3 *Bibliographic Databases for Computer Science*

There are multiple citation indices available at the present [75]. Traditionally, the main source of citation data were Thomson Reuters ISI citation indices, which cover broad range of journals in all major fields of science and humanities. However, Moed [71, p. 119] showed that the coverage of computer science journals is limited. Moreover, the ISI index does not contain many computer science conference proceedings and therefore has limited suitability for the analysis of citation impact in the computer science domain. Therefore, we do not consider the ISI data to be generally suitable for investigation of relations between computer science communities.

In contrast, DBLP [58] provides high-quality manually collected bibliography data covering a substantial range of computer science literature including conference proceedings. As we discuss in more detail in Section 2.4, DBLP has been an important source of data in many studies of information and influence diffusion and dynamics of social communities [30, 51, 80, 15, 24]. For example, by analysing conference proceedings as a proxy of the social structures underpinning the individual fields of computer science, Biryukov and Dong [15] investigated differences and similarities between

computer science communities with respect to their population stability, author productivity, and performance. Despite its high quality, DBLP contains only little citation data. One of the reasons is undoubtedly the fact that the records are inserted manually which makes the citation extraction laborious.

In order to enable large-scale citation indexing of computer science literature, Giles, Bollacker, and Lawrence [38] developed an autonomous citation indexer CiteSeer (now only available in its newer version CiteSeerX [60]) that crawls and extracts metadata, including citations, from the publications on the Web. While the machine extraction of the metadata inevitably has led to errors due to different typographic conventions, CiteSeer corpus contains more extensive citation data [82] than DBLP. Therefore, the citation data from CiteSeer can be used for analytical purposes. For instance, Fiala [32] demonstrated the suitability of CiteSeer data for ranking of computer scientist by their total citation counts.[3] In addition to that, Zhuang et al. [111] used CiteSeer data to measure quality of computer science conferences. The lack of citation coverage of DBLP can thus be at least partially alleviated by its integration with CiteSeer. We shall return to this subject again in Chapter 7.

2.3.4 *Summary*

In summary, patterns of relations between individual researchers or their communities have been illuminated by investigation of citation or co-citation networks. Venues, i. e. conferences and journals, have often been used as a proxy for explicit definition of communities of researchers. However, lack of suitable citation data have prevented investigation of citation impact between computer science conferences. Furthermore, citation data has often been used for evaluational purposes, often resulting in a ranked list of the actors according to various performance metrics. Since the communication practises and research culture in general differ across the different scientific areas, it is generally recommended to use not only one but multiple indicators and analyse only a sample of carefully selected actors that are similar to each other, e. g. within the same field. When comparing across fields, normalised measures that account for the differences between fields should be used. As we discuss in the next section, some of the impact indicators developed by bibliometricians have influenced fundamentally research in other fields such as information retrieval on the Web.

3 A list of ACM SIGMOD E. F. Codd Innovations Award winners was used as a baseline.

2.4 WORLD WIDE WEB

The Web has enabled organisation of and communication between people at an unprecedented scale. The sheer volume of the Web [77] implies that it can be searched only by automated methods. The core problem of information retrieval on the Web is therefore to find the most representative or authoritative web pages. For example, if a user of a search engine like Google queries "world wide web", the user presumably expects an authoritative source on the topic such as a corresponding page on Wikipedia or the home site of the WWW consortium. We present two perhaps most famous algorithms that have been proposed to measure the authoritativeness of web pages: PageRank and HITS. We start with PageRank, because even though it was invented after HITS, its high similarity with the method of Pinsky and Narin [84] makes the transition of our discourse from the citation networks to the Web somewhat easier. Although the two algorithms were originally proposed for the Web, they can be used as a measure of centrality in social networks as well.

The method of Pinsky and Narin is briefly described in Section 2.3.1.

2.4.1 *PageRank*

Similar to the method of ranking high-impact journals, Page and Brin [20] proposed to measure authoritativeness of a web page according to how many links from other authoritative pages it received. Likewise, PageRank score can therefore be understood as a stationary probability of a random walk process on the network of web pages. However, the random walker may reach a "dead end" if it arrives to a web page that does not have any outgoing links. Page and Brin addressed this problem by adding a small probability δ that a walker jumps to a random page anywhere in the network. This guarantees that the walker cannot get trapped in a dead end. Formally, a PageRank score of a node i can be defined as [20], [76, p. 175]:

$$\mathrm{pr}(i) = (1-\delta) \sum_{x \in N_i^{in}} \frac{\mathrm{pr}(x)}{\text{out-degree}(x)} + \delta, \tag{4}$$

where N_i^{in} are the in-neighbours of node i (see Section 2.1). We see that the score of a node is a recursive sum of the scores of its in-neighbours. We also see that the score of a node x is divided equally between all the nodes that x links to. Despite its practical success in information retrieval [20] or social network analysis [24], [55, p. 221], some of the specific differences between citation networks and the Web inspired the development of another popular algorithm—HITS.

2.4.2 *HITS: Hubs and Authorities*

Kleinberg [52] argued that in contrast with citation networks where authoritative sources, e. g. journals, typically endorse each other, some authoritative web sites may purposely neglect other similar web sites. For example, web sites of car manufactures are naturally likely not to link to each other due to competition. In order to measure authoritativeness of web pages, he proposed to distinguish two scores that characterise the position of a page in the link structure of the web: hub and authority scores. The two scores are computed jointly such that a page has a high authority score if it is linked to from many pages with high hub scores. Conversely, a high hub score of a page indicates a high authoritativeness (quality) of the resources it links to. The two scores are therefore in a mutually self-reinforcing relationship. One of the implications is that even if the authorities neglect each other, a high authority of a web page may be induced by an intermediary layer of links from hubs that may be mutually unknown to each other.

It is interesting to observe that such a duality of a node's position may exist also in a social network. Imagine for example a citation network between communities of researchers defined explicitly as sets of researchers that attended the same conference. A link between two communities represents the total number of citations between the papers that were published at the two conferences. Some conferences have very narrow focus, whereas other conferences cover broader range of different areas of one or even more disciplines. We may expect that the papers from the conferences with a broad focus cite many of the papers that were published in the specialised conferences. However, it is less likely that a highly specialised conference would cite many other conference that have narrow focus, because there are typically a few highly specialised venues. Therefore, we may expect that some conferences may act more like a hub, i. e. a place where the researchers aim to disseminate their work across the boundary of their primary field. Similarly, the conferences with narrow focus represent an authoritative venue reflecting the core interests or specialisations of their attendees. We explore this analogy in more detail in the following chapters.

2.4.3 *Summary*

The measures of authoritativeness of web pages HITS and PageRank became widely popular beyond the domain of information retrieval. For example, they were adopted for finding experts in social networks [55, p. 221], or used as a baseline in research of information diffusion [24]. This underlines the wide applicability of network-based measures of centrality and influence

across different areas of science. In the next section, we survey the literature on influence analysis in online discussion communities.

2.5 ONLINE DISCUSSION COMMUNITIES

The Web fundamentally influenced research of social networks and communities because it made many of them *observable*. This enabled investigation of social and information networks that would be otherwise very hard or even impossible to obtain. For the first time in history, phenomena like conversational dynamics [103, 4, 68, 8] information diffusion [3, 10], social influence [93] and community leadership [45] can be traced and studied at a large scale. As in the rest of this chapter, the domains we touch in this section are vast and thus we focus only on the fundamental contributions that are related to our study of cross-community influence.

Since an actor can be a member of multiple communities with a varying degree of belonging, we first discuss the notion of commitment and membership of an actor in a community in Section 2.5.1. After that, in Section 2.5.2 we review the literature related to information diffusion between discussion communities and its maximisation over the individual actors and communities.

2.5.1 *Commitment and Membership of Actors in Communities*

COMMITMENT Online communities in general have been a subject of intensive interest of researchers investigating what factors have an influence on the level of contribution and *commitment* of their members. Kraut and Resnick [53, p. 78] distinguish two types of affective commitment, defined as a "wanting to continue as a member of the group": *identity-based* and *bond-based* commitment. The identity-based commitment arises from the actor's identification with the goals, values, and purpose of the community. The bond-based commitment refers to the attachment of the actor to the community through her ties (bonds) to the other community members. For example, Backstrom et al. [7] observed that an actor is more likely to join a community if many of her friends are already its members. Kraut and Resnick [53, p. 77] argue that actors who are more committed to the community contribute more content, care about the community even if it faces some challenges, and help the newcomers and thus sustain the community in the long term.

Moreover, the highly committed members often play a major role in the conversational dynamics of discussion communities. In a study of USENET discussion communities, Arguello et al. [4] reported that posts that are con-

tributed by more committed, i.e. older, community members are more likely to be replied. Similarly to that, in a study of Yahoo! Groups Backstrom et al. [8] observed that the members who represent a stable core of the community in the long term are 20-times more likely to be replied.

MEMBERSHIP The commitment may be challenging to measure directly because it requires to investigate the motivations of the individual actors. However, the belonging of an actor to a community, i. e. her *membership*, may be defined pragmatically as the level of her activity within the community [68, 80]. In their study of USENET discussion communities, McGlohon and Hurst [68] measured an actor's community membership as the fraction of the actor's overall messages that have been posted in that particular community. Patil et al. [80] took a similar approach and defined membership in scientific communities by distribution of the authors' publications. However, whereas commitment refers to the actor's *motivations* over time, membership corresponds to the actor's *activity* within a certain time frame. Nevertheless, we may say that if an actor maintains high degree of membership in some community over the time, she is highly committed to that community.

2.5.2 *Influence and Information Diffusion*

The fact that an actor may be a member of multiple discussion communities implies that she may share some information in multiple communities simultaneously. For instance, McGlohon and Hurst [68] investigated diffusion of information between USENET discussion communities. A specific feature of USENET is that a message can be sent directly to multiple communities at once, i. e. *cross-posted*. Cross-posting often leads to higher activity within the communities [103, 68]. However, in many discussion communities cross-posting is not allowed. Another approach to analysis of information flow between discussion communities is to model the flow using a network derived from the actors' replies as a proxy [105]. This assumes that the information flows in the opposite direction of a reply, from an actor, who is replied to, towards the replying actor. Figure 9 illustrates the way the information flow is modelled using the replies. We have already encountered with a similar assumption in Section 2.3.1, where Dietz et al. modelled the information flow from cited to citing papers. The rationale in both cases is that a reply (citation) is an explicit engagement with the content of the initial post (paper) and therefore the reply (citation) indicates that the replying actor has ingested the content. Similarly to Dietz et al. who found that information from some of the highly cited papers diffused to other papers, we may expect that information contributed from users of discussion communi-

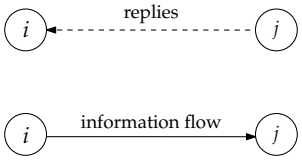


Figure 9: An illustration of the way the information flow is indicated by the replies.

ties who are frequently replied to may diffuse more than information from other users.

In his research of leadership in discussion communities of Google Groups, Huffaker [45] indeed observed correlation between the number of replies an actor received and information flow. He defines the leadership as an ability of an actor "to trigger message replies, spark conversations, and diffuse language". He observed that "online leaders influence others through high communication activity, credibility, network centrality, and the use of affective, assertive, and linguistic diversity in their online messages." Furthermore, he showed that posts contributed by leaders receive a lot of replies by other community members, who also adopt their language, which indicates a diffusion of information from the leaders to the rest of the community. It also suggests that the total number of replies an actor received, i. e. her in-degree in the reply network, is a good indicator of leadership. This could be used potentially for designing efficient communication strategies whereby some information is shared initially with the leaders, who communicate it further to the rest of the community members.

DIFFUSION MAXIMISATION The problem of designing efficient communication strategies has been cast in computational terms as an optimisation problem where the objective is to maximise the spread of information, or generally influence, through a network [51, 57, 23, 19, 43]. The fundamental assumption of diffusion maximisation is that by stimulating—or *targeting*—a small set of *seed actors* embedded in a social network, e.g. by sharing some information with them, or asking them to do some action, the targeted actors trigger a cascade that propagates over as many actors as possible. Diffusion maximisation techniques may enable saving of resources or it may help to avoid information overload. This has important applications in many areas such as public health [100], marketing [3], or innovation management [99]. In public health promotion programs, for example, the diffusion maximisation may be used for efficient information dissemination about diseases and their prevention [100].

DIFFUSION OVER INDIVIDUALS The seed actors are commonly identified by one of the following three ways.

1. The first, heuristic, way leverages some measure of an actor's centrality like in-degree (Equation 1) or PageRank (Equation 4) [51, 24].

2. The second approach is to determine an actor's influence by a simulated process of influence diffusion [51, 23, 57]. See the surveys [92, 43] for an overview of the diffusion models. We further present two models that occur frequently in the literature in Chapter 5.

3. Finally, the third method for identifying seed actors is to measure their influence empirically by investigating who follows whom in their activity [19, 6, 10]. For example, Bakshy et al. [10] conducted a randomised trial in which a population of 253 million Facebook users was divided into two groups: those who were exposed to information that was shared by their friends (neighbours in the network), and those who were not. By comparing the likelihood of some information to occur in the two groups they were able to directly measure the influence a user has on the sharing behaviour of her neighbours as an increase of the likelihood.

DIFFUSION OVER COMMUNITIES In the case of communities like online discussion fora, the scenario is different to information or influence diffusion from a set of seed actors because a stimulus, e. g. a message, is shared with *all* participants in the community. Thus, the problem becomes how to target a stimulus to engage a *set of actors* rather than individuals in a network, such that the stimulus reaches as many actors in the network as possible, i.e. *actor adoption* is maximal. Furthermore, since communities are often centred around specific interests, aims, or affiliations, the problem can be formulated alternatively as a maximisation of the spread of a stimulus across as many *communities* as possible, i.e. *community adoption* is maximal. Maximising adoption over communities may be desirable whenever the stimulus is relevant to a community as a whole. For instance, consider a scenario where a stakeholder of an online discussion system desires the communities to adopt a specific communication practise, e. g. she may desire that the communities do not share or communicate content that is illegal. Even though this may be achieved by a message that would display to every actor, the message may simply be ignored, whereas if they are influenced by their peers, they may be more likely to adopt the new practise. In either case, we refer to the problem of efficient targeting of communities as the *cross-community information diffusion* problem.

The problem of cross-community information diffusion has gained recently more attention of two research groups [30, 69]. First, Eftehkar, Ganjali, and Koudas [30] proposed a model for studying information cascades over communities of researchers extracted from DBLP [58]. The authors assumed that an actor is always a member of all her communities to the same degree. That means that the model they proposed assumes crisp overlapping communities. Second, Mehmood et al. [69] proposed a graph summarisation technique for extracting network of influence relations between implicit communities from the microblogging service Yahoo! Meme. These works have three major limitations:

1. First, they were proposed only for communities with crisp overlap, i. e. they assume that an actor is affiliated to the same degree with all the communities of which she is a member. However, actors frequently belong to a community with a varying degree of membership, i. e. the overlap between the communities is fuzzy [42, 2, 80, 68]. The assumption of crisp overlap is therefore not realistic for those communities. We already argued in Section 2.2.2 on the example from Figure 6b that the way the communities overlap may fundamentally affect the distribution of influence between the communities.

2. Second limitation is that these methods assume a static network leaving unanswered the question of their performance when used on dynamic graphs.

3. And finally the third limitation is that the authors evaluated the performance of their models only with respect to the overall adoption by the actors at the end of the diffusion process. It is thus not clear what is the behaviour of the proposed models with respect to community adoption.

Therefore, to the best of our knowledge, a comprehensive approach addressing these limitations is lacking.

2.5.3 *Summary*

To sum it up, membership of individual actors in the communities can be quantified as the distribution of their activity over the communities. Their conversational activity can also be leveraged for modelling information flow. One of the ways how to model the flow is to construct an information flow network from the conversational interactions between the actor. Since some of the users of online communities were observed to be more influential

than others, the information flow may be efficiently maximised by engaging with the *influential actors* only. However, in discussion communities the information is shared frequently with the whole community and not with individuals and thus the problem becomes to maximise cross-community information diffusion by targeting *influential communities*. Although this problem has gained recently some attention, to the best of our knowledge there is no comprehensive approach that addresses this problem for fuzzy overlapping communities in dynamic networks; and that maximises the information diffusion over both the individual actors as well as over their communities.

2.6 CONCLUSION AND LIMITATIONS OF THE STATE-OF-THE-ART

We have surveyed several fundamental contributions in studies of impact and influence in social networks, information networks, and in online discussion communities. There is comparatively little work on how communities as entities influence or are influenced by each other. Moreover, the few methods that were proposed to measure influence or centrality of a community have several *limitations*:

L1 They characterise the relation between a community and the rest of the individual actors [31].

L2 They assume the communities do not overlap at all [34], or share their members equally, i. e. to have a crisp overlap [69, 30].

L3 They were developed and evaluated only for static networks and communities. Therefore, they did not provide any insights into cross-community influence over time.

L4 The two recent models that were proposed to find influential communities [69, 30] were evaluated only in the context of maximising the information diffusion over either microblogging or scientific communities. The efficacy of either of the two methods beyond the domain they were evaluated in is not known.

L5 None of the discussed community-level influence measures integrates any additional information, e. g. extracted from the textual data.

Therefore, to the best of our knowledge, there is no extensible model that enables measurement, analysis, and exploitation of dynamic cross-community influence in multiple classes of social communities, such as discussion or scientific communities, leveraging both structural and text information.

The aim of this thesis is to address those limitations. By overcoming all these limitation, we expect to provide a crucial contribution towards the

general understanding of cross-community influence and the resulting effects. In the next chapter, we present the core of our framework for cross-community influence that we call COIN (limitations L1 and L2). The core framework is based purely on structural features. We evaluate the core framework in analysis of cross-community influence in two types of online discussion communities in Chapter 4 and Chapter 5 (L3 and L4). After that, we generalise the framework in Chapter 6 in order to investigate the topics that underpin the cross-community influence relations (L5). Finally, in Chapter 7, we demonstrate the COIN framework in a study of cross-community influence between communities of computer science researchers (L3 and L4).

3

COIN: CROSS-COMMUNITY INFLUENCE ANALYSIS FRAMEWORK

In Chapter 2 we argued that the methods that have been previously proposed in the literature to quantify cross-community influence have several limitations. The fundamental limitations of the previous methods arise from the way they represent the communities and their interactions. For example, the previous methods assumed only non-overlapping or crisp-overlapping representations that are less realistic than the fuzzy representation for communities like online discussion or scientific communities [68, 80]. Furthermore, each of the previous methods was evaluated on only one type of a social system [30, 69, 34] and they often addressed only a particular application domain, e. g. information diffusion [30, 69]. Therefore, their applicability or extensibility beyond that domain is not known. As a result, the efficacy of the previous methods for the measurement, analysis, and exploitation of cross-community influence in different classes of dynamic social communities is limited. For example, we show later in this chapter that influence between many discussion communities cannot be measured using any of the previous methods that use the crisp representation of communities.

The difference between crisp and fuzzy overlaps is illustrated in Figure 4 on page 16.

We take the first step towards addressing these limitations. We present the core of our extensible computational framework for cross-community influence—COIN. The core model is the essential building block on which we will base further extensions. In the next section, we present the assumptions of the core framework behind the data representation. After that, in Section 3.2, we formulate the main hypothesis of our structural approach for the measurement of cross-community influence. We develop the hypothesis into a set of measures that constitute the core of our framework in Section 3.3. Finally, in Section 3.4 we discuss the limitations of the core framework and how we address them in the following chapters.

3.1 REPRESENTING DATA IN THE CORE COIN

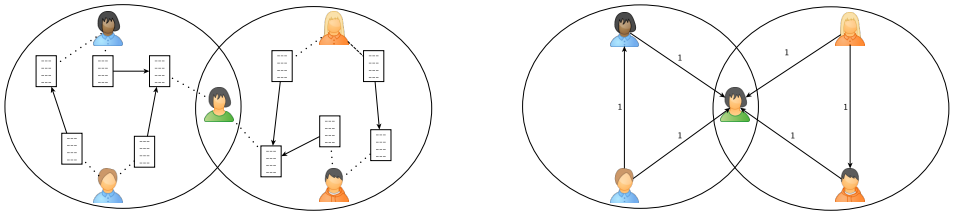
Before we develop the core measures of COIN, we present the way we represent data. As we stated in Section 1.1 on page 5, our focus in this thesis are social communities of actors who interact by responses in a form of e. g. posts in discussion communities or papers in scientific communities. In general, we refer to any information artifact through which the actors

interact as *document*. Therefore, the actors author documents that can be in response to one another. We may say that the author of the responded document *influenced* the responding author towards *activity*—she *stimulated* the response. Therefore, we adopt an *activity-based* notion of influence.

For the sake of clarity, we first assume the simplest model of interactions in which:

1. a document is authored by exactly one author;
2. the responses occur only between documents from the same community;
3. each document exclusively belongs to a particular community.

This corresponds to a broad spectrum of online discussion communities where a post (document) is always contributed to a particular community and other posts may be in response only to the posts from the same community. However, as we mentioned in Section 2.5.2 on page 31, this is not the case in USENET, where a message may be cross-posted to multiple communities. Likewise, in scientific communities a research paper is often co-authored by a team and the papers may cite (respond to) papers from other communities. We return to these more general cases at the end of the chapter.



(a) responses between documents

(b) responses between actors

Figure 10: An illustration of the simplest representation of interactions, where a response always belongs to a particular community. A directed link in Figure 10a depicts that a document at the source of the link is in *response* to the document at the sink of the link. An authorship of a document is denoted by an undirected dotted link.

An example of the simplest model of interactions is illustrated in Figure 10. We see in Figure 10a two communities that share one (green) actor, but they do *not* share any documents nor does any document respond to another document from a different community. Since the responses in Figure 10a are between documents and not their authors, the figure illustrates

an *information network* of relations between documents. In order to analyse social interaction among the actors, we can derive a *social network* depicted in 10b from the information network by propagating the responses between the documents to their authors. That is, we connect actor i to actor j if any document authored by i was a response to any document authored by actor j.

Throughout the thesis, we represent the social network as a dynamic directed graph. In order to analyse the dynamic changes in the network, we propose to segment the data using a sliding time window. As a result, each segment can be represented as a *snapshot* or *time-slice* graph $G = (V, E)$ consisting of $n = |V|$ actors participating in k communities. The ties in the social network can be weighted, e. g. by the number of responses between the actors. For example, we see that there is one response between each pair of actors connected by a tie in Figure 10b. We represent the ties E and their weights as a weight matrix $\mathbf{W}$.

We describe our elementary terminology in more detail in Section 2.1 on page 12.

Having the actors' interactions represented as a network, we can measure their influence using a centrality measure like in-degree (Equation 1 on page 18). An actor with a high in-degree highly *stimulates* the other actors—she triggers responses from them. As we argued in Chapter 2, this can be interpreted as a measure of influence. In Figure 10b, the actor with the highest in-degree overall (in total 4 responses) is the green actor at the intersection of the two communities.

The green actor also has the highest in-degree within each of the communities. We see that in total the actor stimulated 2 responses within each community. That is, her in-degree in the confined subgraph (see Definition 5 on page 16) of each of the communities is 2. We also see that apart from the green actor at the intersection there are no responses between the actors of the two communities. This is a consequence of our second assumption that the responses always occur only between documents of the same community. In cases like that, it is the actors at the intersection through which the communities interact.

The fact that the green actor has the highest centrality and is the only point of interaction between the two communities suggests that the distribution of her membership over the two communities also determines the distribution of influence between the two communities. If, for example, the green actor belongs predominantly to the left community, it suggests that the community on the right is strongly influenced by the left community. This is because the most central member of the right community belongs mostly to the left community. Naturally, this is not possible to represent by non- or crisp-overlapping communities, because those representations do not allow the actors to have different membership levels in several communities. In

the next section we describe how the overlap between the communities and the distribution of the membership of the individual actors fundamentally affects the cross-community influence relations.

3.2 THE HYPOTHESIS OF STRUCTURAL CROSS-COMMUNITY INFLUENCE

The level of actors' membership in their communities may differ for each actor and community. While one community may consist entirely of actors with high membership in the community, another community may have an actor base with more heterogeneous distribution of memberships. Later in the section we argue that the differences in memberships of the actors fundamentally shape the relation between any two interacting communities. For that reason, we propose to differentiate actors according to their level of membership in a community into *focal* and *alter* members of the community.

Proposition 1 *Let u and v to be overlapping fuzzy communities according to Definition 3 on page 14. Membership of each user in each community is defined by an activity level in each community. Following Definition 3, we represent the memberships in communities u and v by the membership functions φ_u and φ_v that map a member to her membership in the community. For any actor i whose membership in community u is higher than in v, i.e. $\varphi_u(i) > \varphi_v(i)$, we propose to call:*

- *actor i a* focal member *of community u with respect to community v, and* alter member *of community v with respect to community u;*
- *community u the* focal community *of actor i with respect to community v;*
- *community v an* alter community *of actor i with respect to community u.*

We define the set of actors for whom u is the focal community as the focal members *of community u with respect to v. Conversely, we define the set of actors for whom u is an alter community as the* alter members *of community u with respect to v.*

In order to simplify the language, we state further in the text only that a user/community is focal/alter without specifying with respect to what community. It should be clear from the context with respect to what community the classification was done.

We have already argued in Section 2.5.1 on page 30 that the membership of an actor can be measured as a distribution of her activity. In the case of communities of researchers where venues (journals or conferences) are used as a proxy for communities, the membership can be defined as a distribution of the author's publications over the venues (communities) [80]. Similarly,

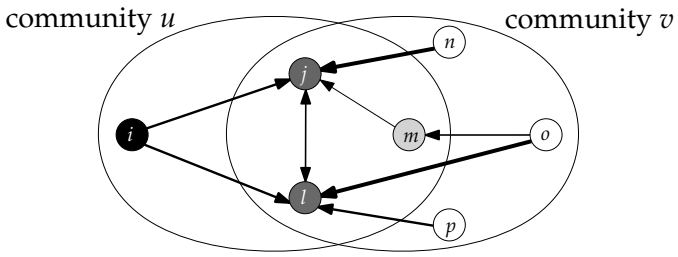


Figure 11: Example of impact from community u to v. Nodes are actors connected by links whose thickness reflects the number of responses. The shading expresses community affiliations, such that the darker (lighter) the node is, the more is the actor member of community u (v).

the membership of users of online discussion communities can be quantified as the distribution of their messages over the communities [68].

As actors have quantitatively different memberships between communities, this results in asymmetric relations between communities. If focal members of community u are highly influential within community v, then we may say that community u influences v. If the influence is measured using a centrality score in a response network, the influenced community v depends on the focal members of community u to stimulate activity in v. In other words, the focal members of community v are *not* the most influential within their own community. This leads us to the following hypothesis:

Hypothesis 1 *Given two fuzzy communities u and v defined according to Definition 3 on page 14, the influence of community u on v can be quantified as an increasing function of membership of the actors from community u, φ_u, and their centrality in community v.*

As an idealised example, consider two communities u and v as depicted in Figure 11, in which the nodes represent actors connected by their responses, e. g. replies in the discussion communities. The link thickness reflects the tie strength, e.g. in a discussion community it can represent a number of replies from one actor to another. The shading of nodes shows an actor's community membership: the darker the node the more the actor belongs to community u; and the lighter it is, the more the actor belongs to community v. We see that actors $\{n, o, p\}$ from v maintain strong ties to the actors $\{j, l\}$ whose focal community is u. Therefore, while the actors $\{i, j, l\}$ tend to maintain strong ties with each other, they are also *highly central* within community v, which is their alter community. According to our hypothesis even though community v has more members than u, community u has a high influence on v because the most central actors of v belong more to community u.

In order to evaluate Hypothesis 1 and explore the cross-community influence in the systems that we have analysed, we developed a computational model, whose main concepts are introduced formally in the next section.

3.3 THE CORE MEASURES OF COIN

Our hypothesis requires us to examine two aspects of the interactions between communities. First, since the communities are fuzzy overlapping, it requires us to take into account the differing memberships of the actors in each community. Second, it requires us to quantify the tendency of an actor to stimulate the members of a community. To illustrate this intuition, recall the idealised scenario from Figure 11: focal members of community v, $\{m, n, o, p\}$, respond frequently to its central actors $\{j, l\}$ whose focal community is u, and therefore u has an influence on v. Following these intuitions we define COIN formally here. First, we define a measure of influence between pairs of communities. After that, we derive a set of aggregated measures that quantify to what extent a community influences or is influenced by the other communities. Please recall that our data represents n actors and k communities.

Let us define an $n \times k$ **membership** matrix $\mathbf{M} : \mathbf{M}_{iu} \in [0, 1], \forall\, i : \sum_{x=1}^{k} \mathbf{M}_{ix} = 1$ that represents the affiliations of n actors among k communities. The column $\mathbf{M}_{\cdot u}$ represents the fuzzy set C_u of the members of community u (see Definition 3 on page 14). Since the rows of the matrix are normalised, each actor is "divided" into her communities and the sum over all the memberships is equal to the total number of actors, i. e. $\sum_{x=1}^{n} \sum_{y=1}^{k} \mathbf{M}_{xy} = n$. The normalisation of the rows also makes the memberships of the individual actors comparable. The matrix $\mathbf{M}$ can be known a priori from a field survey, it can be derived from activity traces of the actors, or determined by a community detection algorithm [33, 42, 106]. In our analysis we set $\mathbf{M}_{iu} = |P_{iu}| / \sum_{x=1}^{k} |P_{ix}|$, where P_{iu} is the set of documents actor i contributed to community u. Hence we adopt the activity-based notion of community membership that was presented in Section 2.5.1 on page 30.

An influence of any given actor within her communities can be formalised as an $n \times k$ **centrality** matrix $\mathbf{C}$ with elements $\mathbf{C}_{iu}$ representing the influence of actor i on the other actors of community u. Depending on the nature of the relation between the actors, the centrality can be obtained by an actor centrality measure on a confined subgraph of the community (see Definition 5 on page 16), e. g. in- or out-degree (Equations 1 and 2). Alternatively, it can be provided explicitly, for example as points or badges awarded to the actor. In the next chapters, we make use of this flexibility in the definition of $\mathbf{C}$ and measure it differently for each of the systems that we have analysed.

Following our Hypothesis 1 on page 41 we define the function that measures the cross-community influence as a *product* of the membership and centrality matrices:

Definition 6 *We define $k \times k$ cross-community* impact matrix $\mathbf{J}$ *as a product:* $\mathbf{J} = \mathbf{S}^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{C}$*, where* $\mathbf{S}$ *is a $k \times k$ diagonal matrix with vector of community sizes* $\mathbf{s}$ *on its diagonal. The* impact $\mathbf{J}_{uv}$ *of community u on community v is therefore defined as the normalised sum of centralities of the members of u within community v, weighted by their membership in u:*

$$\mathbf{J}_{uv} = \frac{\sum_{x=1}^{n}(\mathbf{M}_{xu} \cdot \mathbf{C}_{xv})}{\mathbf{s}_u} \tag{5}$$

Sizes of social communities often vary [79], and thus, for example, a very big community can, from its raw size, accumulate high impact despite the low membership of its members. The factor $\mathbf{s}_u$ in the divisor therefore guarantees that the impact of community u remains unbiased by its size. As we model the communities as fuzzy sets (Definition 3 on page 14), we define the size of community u as the *cardinality* of the set of its members C_u. The cardinality of the fuzzy set C_u can be naturally defined as the sum of the memberships of its elements, i. e. $\sum_{x=1}^{n} \mathbf{M}_{xu}$ [109].

However, since the membership of an actor may vary within $[0, 1]$, the cardinality of the set C_u may be lower than 1. This happens for instance if community u consists of only a few members with very low membership, i. e. $\forall\, i : \mathbf{M}_{iu} \ll 1$. In such cases, the numerator in Equation 5 would be divided by a number lower than 1, which would inflate the value of the impact and thus bias it towards small communities. We address this issue by introducing a community size threshold θ below which we do not normalise the impact by the size. In our experiments, we used $\theta = 1$. Therefore, we normalise the impact of the community u by the size that is at least 1, i. e. we set $\mathbf{s}_u = \max(\sum_{x=1}^{n} \mathbf{M}_{xu}, 1)$.

Each row $\mathbf{J}_{u\cdot}$ of the impact matrix contains the impact the community u has on each other community, including the community u itself. We call this self-impact $\mathbf{J}_{uu}$ an *independence* of the community, because it measures to what extent the focal members of community u are also central in it:

Definition 7 *We define the* independence *of community u as the self-impact of the community* $\mathbf{J}_{uu}$*.*

AGGREGATE MEASURES While the distribution of the impact values over one row of $\mathbf{J}$ is useful for low-level cross-community analysis, the sum over the row represents the *overall* impact the community has on others. We define this as the *importance* of the community. In order to emphasise the cross-

community impact, it is useful to exclude the independence from the overall impact. This leads us to the following definition:

Definition 8 *We define the* importance *of the community as a sum over the row without the diagonal element representing its independence. Formally, the vector of importance values can be computed as:*

$$\mathrm{imp}(\mathbf{J}) = \mathbf{J}\mathbf{1} - diag(\mathbf{J}), \tag{6}$$

where $\mathbf{1}$ *is a column vector of ones of length k.*

Whereas the rows of the impact matrix $\mathbf{J}$ represent the impact each community has on others, each of the *columns* represent the distribution of impacts other communities have on the community. In particular, if many values of the column are higher than the independence of the community, it indicates that many of the most central actors in the community are its alter members. Such community therefore *depends* on other communities, because its activity is driven by actors whose main interests reside somewhere else. Hence the definition of the community's dependence is:

Definition 9 *We define the community's* dependence *as the sum of the impact other communities have on it. Formally, the vector of dependence values is:*

$$\mathrm{dep}(\mathbf{J}) = \mathbf{J}^{\mathrm{T}}\mathbf{1} - diag(\mathbf{J}) \tag{7}$$

It is possible that the impact between two communities u and v is mutual, i. e. $\mathbf{J}_{uv} > 0$ and $\mathbf{J}_{vu} > 0$, but a significant difference between the impacts, e. g. $\mathbf{J}_{uv} \gg \mathbf{J}_{vu}$, suggests that the communities are in an asymmetric relationship. For instance, if community u has a high impact on community v, as idealised in Figure 11, then we find that the most central actors of v are focal members of community u, rather than v. Using the introduced concepts we formalise this intuition by the following definition:

Definition 10 *We say that an impact of u to v is* strong *if it is at least as high as the independence of v. Otherwise we say that the impact is* weak.

While some communities may impact a relatively small circle of other communities, others may be broadly influential. For instance, a community of system administrators may have an impact on the whole system. Analogously, a community may be influenced by many other communities or it may be strongly influenced just by few communities. An analysis of the distribution of importance (rows of the impact matrix) and dependence (columns) gives a clear indication of whether a community's influence/dependence is largely dispersed or narrowly focused. We quantify the

heterogeneity of importance or dependence as an entropy of a row or a column of the impact matrix $\mathbf{J}$. Because some elements of $\mathbf{J}$ may be 0, let us use the convention $\log_2(0) = 0$. Furthermore it is necessary to normalise the rows of the matrix in order to obtain probability distributions of impact, i. e. $\mathbf{J}^N_{uv} = \mathbf{J}_{uv} / \sum_{x=1}^{k} \mathbf{J}_{ux}$. Formally:

Definition 11 *The normalised* importance entropy *of community u is defined as*

$$\mathrm{ent}_{\mathrm{imp}}(u, \mathbf{J}) = -\frac{\sum_{x=1}^{k} \mathbf{J}^N_{ux} \log_2 \mathbf{J}^N_{ux}}{\log_2 k} \tag{8}$$

The normalised dependence entropy $\mathrm{ent}_{\mathrm{dep}}(u, \mathbf{J})$ *is defined similarly but on the transpose* $\mathbf{J}^{\mathrm{T}}$*:*

$$\mathrm{ent}_{\mathrm{dep}}(u, \mathbf{J}) = \mathrm{ent}_{\mathrm{imp}}(u, \mathbf{J}^{\mathrm{T}}) \tag{9}$$

Both measures range within $[0, 1]$. The more the importance (dependence) of community u is equally distributed, the more the entropy value is close to 1. We note that in the case of entropy we *include* the diagonal elements (independences), because we wish to differentiate whether the most of the community's total impact is concentrated *within* that community or not.

HUB COMMUNITIES Our definitions of *importance* and *dependence* has a parallel with the *authority* and *hub* scores as quantified by the HITS algorithm for authority measurement of web pages that we discussed in Section 2.4.2 on page 29. They are both underpinned by a similar intuition of the differences between incoming and outcoming links. Recall that HITS assigns high authority score to pages that are frequently pointed to by the hubs and thus the authority score is an indicator of the quality of the links *incoming* to the page. For the sake of simplicity, let us assume that the communities represent users of discussion fora and that an actor's centrality in a community is measured by her in-degree, i. e. it expresses the ability of the actor to stimulate conversations in the community. Therefore, a high importance of a community indicates high number of links (replies) *incoming* to the community's focal members from the members of other communities. Furthermore, HITS assigns high hub scores to the pages that frequently point to other pages and thus it indicates the quality of its *outcoming* links. Analogously, a high dependence of a community means that the focal members of the community frequently respond to members of other communities. In other words, the dependence indicates the number of *outcoming* links from the focal member of the community to the members of other communities. We therefore expect that a community that is highly dependent and whose dependence is highly distributed (i. e. high dependence entropy) acts as a *hub*:

Proposition 2 *A community that is highly dependent on many other communities is likely serving as a common meeting place that brings together focal members of other communities—the* hub *of the system.*

Therefore, we expect that if there are any hubs in the system, they can be identified by a high dependence and entropy.

However, we note that HITS does not allow us to address our research questions (see Section 1.1 on page 5), because both hub and authority scores characterise the relation of a node with the rest of the network. Therefore, it is unsuitable for investigation of pair-wise influence relations between communities. Conversely, our framework contains a set of measures that characterise both pair-wise relations between communities and aggregate measures that characterise the relation of a community with all the other communities.

3.4 SUMMARY AND LIMITATIONS OF THE CORE FRAMEWORK

We have developed our main Hypothesis 1 into the simplest version—the core—of our cross-community influence framework, COIN. The fundamental concept of COIN is *cross-community impact* that quantifies to what extent one community influences another community. The impacts between any pair of the communities are represented by a *cross-community impact matrix* or simply *impact matrix* $\mathbf{J}$. We call the self-impact of a community its *independence*. We derived several aggregate measures of impact. The *importance* of a community is the total cross-community impact the community has on other communities. Conversely, the *dependence* of a community quantifies the total impact *other* communities has on the community. Furthermore, we proposed to measure the heterogeneity of the importance and dependence by entropy. We explained how a community that is highly dependent on many other communities acts as a *hub*. Finally, if the cross-community impact of community u on community v is at least as high as the independence of v, we say that the impact of u on v is *strong*. Otherwise we say that the impact is *weak*.

Apart from the definition of membership as a distribution of an actor's activity over the communities, there are other possibilities that may provide alternative or complementary picture of the cross-community relations. For example, since the level of overall activity may differ across the actors, we may want to compare the actors based on how active they are in the system. One approach is to use the most active actor to benchmark the activity of others. The individual memberships would then be normalised by the total number of posts contributed by the most active actor. This means that a high membership would indicate a high activity relative to the maximal activity

of *any* actor in the system. Membership measured this way should capture differences in actors' overall commitment to the system. However, the sum over the memberships of an actor would not generally sum to 1. This issue could be addressed by introducing a "surrogate" community that would represent the differences in activity between the most active actor and the other actors. Assuming that the most active actor spends all her time in the system, the surrogate community could be interpreted as the representation of the out-of-system activity of each actor. Although this alternative measure of membership may provide interesting and complementary insights into the cross-community relations, we leave it for future work. In the following, we use only the definition of membership as a distribution of the actor's activity, because it does not require the surrogate community and it has been already used in the literature [80, 68].

Even though the key notions of focal and alter members, and cross-community impact are all defined on the level of pairs of communities, they can be applied in an analysis of more than two communities. We propose two approaches to the analysis of multiple communities: thresholding and aggregation. First, the notion of strong impact allows us to reveal only the significant relationships between individual communities, because it corresponds to impacts that are higher than the self-impact of the influenced community. Second, the aggregate measures characterise the overall relationships a community has with all the other communities. As we show in the following chapters, the aggregate measures enable insights into what position a community has among many other communities.

The core COIN forms a fundamental building block that may be further extended by supporting various operations on communities and their impacts. For example, more research is needed to explore what effect on impact or influence/dependence we may expect if a community merges with another community, or, conversely, if a community splits into two or more descendant communities. Therefore, the core measures may be developed into an algebra with operations like addition (merger) or division (split). Such extension of our model may indeed be very useful for a practical manipulation of impact with respect to changes of communities. However, we leave this extension for future work and we investigate the dynamics of impact by applying sliding time-window on the data as described in Section 3.1.

The core COIN is based on several assumptions about the data. The model introduced in this chapter is compatible with a broad range of discussion communities. For the sake of clarity, we assumed that a document:

1. is always authored by exactly one actor;

2. may be in response to another document, but both documents must belong to the same community;

3. belongs exclusively to one community.

In the following, we discuss several possible ways to ease the limitations implied by these assumptions.

A document can be frequently co-authored, e. g. a research paper is often written by a team of scientists. There are multiple ways how to deal with this problem [71, p. 273]. One way is to assign the document to each of the co-authors. This in effect creates multiple copies of the document, each copy uniquely belonging to each of the co-authors. The social network of responses is then obtained the same way as described in Section 3.1. Furthermore, in scientific communities defined using venues as a proxy, a paper often cites (responds to) a document from another community. We return to these subjects in Chapter 7.

In some cases a document may belong to multiple communities. One example is USENET, where a message can be cross-posted to multiple communities. In order to determine to which community a cross-posted message belongs, McGlohon and Hurst [68] proposed to measure the ownership of a message by a community by means of the membership of the author within the community. Therefore, if an actor cross-posts a message to two communities u and v, but if her membership is much higher in u than in community v, then also the cross-posted message belongs more to u than to v.

Finally, we did not consider any additional information that may be available in the documents, e. g. extracted keywords. Keywords may provide valuable insights into *topics* that are associated with the cross-community influence relations. For example, we may want to investigate which communities were highly influential with respect to a particular topic like "music" or "computer games". We generalise COIN in order to capture topics in Chapter 6.

However, the textual content of the documents may not be available due to, for instance, privacy or technical reasons. Therefore, the purely *structural* COIN that we have developed thus far still promises to deliver many insights where the additional textual information is not available. Indeed, we use the structural COIN to reveal influence relations between discussion communities and their changes over the time in Chapter 4. We also demonstrate how to leverage the structural model for efficient information flow in Chapter 5.

4 CROSS-COMMUNITY INFLUENCE IN DISCUSSION FORA

In the previous chapter, we introduced the hypothesis of structural cross-community influence and developed it into a core of the computation model that we call COIN. In this chapter, we evaluate COIN on data from two different online discussion systems. While the first, Boards.ie [17], is a general-purpose online discussion board and as such is the largest website of its type in Ireland, the other, SAP SCN [88], is business-driven technical support fora run mainly to provide question-answering facility to customers of SAP. In both cases we define a community as a set of users that participate in a forum, i. e. the fora serve as a proxy for the community structure. We use the synonymous terms "user" and "actor", and "forum" and "community" interchangeably.

We used COIN in order to validate Hypothesis 1 on page 41 by an exploratory and qualitative analysis of Boards.ie and SAP SCN. First, we describe the data that we analyse in Section 4.1. In Section 4.2 we analyse influence between pairs of communities by investigating strong cross-community impact. The results show that in many cases the most influential communities, e. g. the moderating and administrating communities, are relatively small and often private communities whose members are significantly more capable to disseminate information than the other actors. Opposite to that, the busiest and largest communities are often strongly influenced by the other communities. After that, in Section 4.3 we move our focus to the aggregated level of overall impact a community has had over time; or, analogously, overall impact other communities have had on the community. We demonstrate that the aggregate level is useful for discovery of global trends of cross-community influence in the system, e.g. an increasing diversification of influence with the rising size of the system and an emergence of globally influential communities. We end with the discussion of the main findings of this chapter in Section 4.4.

Recall that the impact is strong *if it is at least as high as the independence of the influenced community (Definition 10 on page 44).*

4.1 DISCUSSION FORA DATA

Both Boards.ie and SAP SCN, "Boards" and "SAP" hereafter, are structured according to themes into *fora*, optionally further into their subfora, and finally into *threads* of *posts* centred around a particular conversation topic. Each post has an author, who can be either a registered *user* or a guest.

Since all the guests' posts in Boards are stored with the same user identifier, we omitted them from the analysis. There are no guest users in SAP. A set of users who have posted at least once to any forum within a certain time-period constitute a *community* of that forum in the period. Even though there is no direct way to post a message into multiple fora (i.e. to cross-post it), the users can and do participate in multiple fora. Threads have a tree-like structure as one post can be in *reply to* another one.

A reply in Boards has a different motivation than a reply in SAP. Due to its question-answering character, the reply activity of SAP users is driven by their aim to solve a problem. There is a reward system based on points the users can award to the most helpful answers. In contrast with that, Boards fora serve more as a place for socialising, information sharing, and entertainment. The flexibility of COIN allows us to take this distinction into account by defining centrality differently for each system. While in Boards we want to focus our analysis on communities that stimulate activity in general, in SAP our aim is to uncover communities whose kernel of expert users are crucial for problem solving in other communities.

The set of users tied by the who-replies-to-whom relation forms a directed dynamic *reply network*, as the reply ties change in time. In the case of Boards, the ties are weighted by the number of replies from one user to another within a given time period. We set the centrality $\mathbf{C}_{iu}$ of user i within community u as the in-degree (Equation 1 on page 18) of user i in the confined subgraph (Definition 5 on page 16) of community u, i.e. to the total number of replies user i received to the posts she contributed to forum u. We chose in-degree for our experiments because the reply behaviour is the cornerstone of the conversational dynamics; it is a well-established heuristic for influence maximisation [51, 92]; it was found to correlate with actor leadership in online discussion communities [45]; and it has a clear interpretation. A user is therefore highly central if she stimulates high activity in the community, i.e. receives many replies.

Since a reply in SAP can be awarded points by other users who found the reply useful, we can measure a user's influence by the number of points she has received for her answers. Therefore, we set the weight of a tie (i, j) in the SAP reply network to the total sum of points that the replies from user i to j received. The weight thus measures how much the users in total valued the answers from i to j. The centrality $\mathbf{C}_{iu}$ is then defined as the out-degree (Equation 2 on page 18) of actor i in the confined subgraph of community u, i.e. the total number of points user i received for the replies (answers) she has contributed to community u. Unlike in the case of Boards, where a user is highly central if she receives many replies, we used out-degree

because this allows us to leverage the explicit information of the users' influence measured by the points.

Since the mutual dynamics of communities can be highly volatile in time, we segment the data using a sliding time-window and analyse the changes between the subsequent snapshots of the resulting sequence. Table 1 presents some basic statistics of the analysed data. Please notice that there are many more ties per user in Boards than in SAP. This suggests that the behaviour of Boards users is more conversational than that of SAP users.

	Boards	SAP
# snapshots	448	41
# communities	636	33
mean # of users per snapshot	2,093	1,567
mean # of ties per snapshot	9,656	4,423
ties per actor (over all snaphots)	59	6
# post	8,189,148	420,369
# reply	7,524,427	321,471

Table 1: Elementary statistics of the analysed data-sets. The hash symbol (#) means "a number of".

TIME-WINDOW SELECTION The choice of time-window length clearly affects the results of the analysis. Since our notion of impact is based on the activity in the overlap of the communities (see Chapter 3), the window length should capture as much of that activity as possible, yet still be fine enough to uncover changes in users' behaviour. Let $\tau(x)$ be a *minimum* time it took an author of post x to contribute a message to another forum, i.e. a *cross-fora posting waiting time*. If the author has not posted to any other fora, then $\tau(x) = \infty$.

In order to find a suitable time-window size, we sampled 10,000 posts and investigated the distribution of $\tau(\cdot)$. Table 2 lists values of the empirical distribution function of $\tau(\cdot)$ for some selected times. It turned out that for approximately 85% of the postings to Boards a user has posted again into another forum within 7 days; by doubling the window size to 14 days we found that the cross-posting activity only increased by 4% to 89%. Only in 3% of the cases did a user not post to any other forum. Therefore we chose a one-week window for our analysis of Boards cross-forum activity.

Similarly, we found that for approximately 49% of the postings to SAP a user has posted again into another forum within 60 days, while by dou-

bling the window size the increase of cross-posting activity is again only 4%. Therefore we chose two-months window for the analysis. In contrast with Boards, we observed a much lower level of cross-fora posting activity in SAP—in 40% of the postings a user has not posted to any other forum whatsoever. We believe that the reason may be the difference in their utility. While Boards is primarily a place for socialising and discussion of a broad range of topics, the users of SAP may be more focused on a particular topic related to their expertise or problem. A similar distinction between social and technical fora was also observed by Shi et al. [90].

t (days)	1	7	14	60	120
Boards	0.69	0.85	0.89	0.94	0.96
SAP	0.26	0.35	0.39	0.49	0.53

Table 2: Values of the empirical cumulative distribution function of $\tau(\cdot)$ for selected waiting times.

In total the Boards data was segmented into 448 weekly snapshots between Monday 12. 7. 1999 and Sunday 10. 2. 2008, and SAP data into 41 bi-monthly snapshots between 1. 5. 2004 and 28. 2. 2011.

4.2 PAIR-WISE INFLUENCE ANALYSIS

Both SAP and Boards data-sets span years of existence of the system, which makes the analysis of all possible $k(k-1)$ pairs of cross-community impact impractical. However, it is possible to narrow down the analysis by investigating *strong* impact only, i.e. only the impact that is at least as high as the independence of the influenced community (see Definition 10 on page 44). For example, this reduces the 3,251,098 impacts between two distinct Boards communities in total over all the time snapshots by more than 300 times into 9,856 strong impacts. In the next section, we further analyse the weak impact by utilising the aggregated measures of community importance and dependence.

4.2.1 *Influence Between Pairs of Boards Communities*

Due to space reasons, we illustrate the benefits of pair-wise impact analysis on most frequently observed strong impacts. Table 3 lists the five most frequently observed strong impacts between Boards fora.

community u	# users in u	community v	# users in v	SI
Moderators	53	Reported Posts	116	32
FNWAI	5	Poker	134.96	31
The Thunderdome	31	After Hours	339	27
PI Mods	4	Personal Issues	197	24
Administrators	4	Feedback	90	20

Table 3: Five most frequent strong impacts SI of Boards community u on community v. We present the number of users in each community averaged over all the time snapshots and rounded to integers for the sake of brevity. The hash symbol (#) means "a number of".

Influence of Moderators *and* Administrators

In 32 distinct weeks, the impact of Moderators on Reported Posts was strong, which we found intuitive, since Moderators is the community of users whose role is to facilitate the discussions in other communities and Reported Posts is the place where the users of Boards may report any misconduct in other fora. A similar relationship is between Administrators and Feedback, because Feedback is a place where the users express their opinion about the system.

Influence of FNWAI *on* Poker

While this could be easily guessed solely by the names of the fora, the impact from FNWAI (Fold, No, Wait, All In) on Poker is not obvious. Manual inspection of the content of FNWAI revealed it was a small community of elite poker players. COIN thus enables identification of cross-community impact that may be otherwise left unnoticed.

Influence of The Thunderdome *on* After Hours

After Hours, the most active and biggest forum according to Table 4, is a general meeting space, without a particular topic, visited by its members in order to chat [1]. It seems, however, that many of its highly central members belong to other, more focused, communities, as we observed in total 487 strong impacts by 147 distinct communities (the highest number of all the fora). One of them, The Thunderdome, is focused on mutual provocation and insulting among its members under the agreed rules [96]. Since the impact measures the ability of one community to *stimulate* another one, it is natural that a community specifically focusing on provocation was recognised as influential.

Influence of PI MODS *on* PERSONAL ISSUES

PERSONAL ISSUES is arguably of high importance for many users because it offers them a discreet opportunity to seek advice or help in many difficult real-life situations like personal relationships. Clearly, such discussion needs to be protected from unhelpful comments, and thus it is moderated by a dedicated moderating community—PI MODS.

community	in-degree	# user	# post
AFTER HOURS	72	339	1472
THE CUCKOO'S NEST	74	90	930
POKER	39	135	903
BEER GUTS & RECEDING H.	86	99	860
SOCCER	68	145	859

Table 4: The average group in-degree, number of active users, and number of posts for the 5 communities with the highest average post counts. The values are rounded to integers for the sake of brevity. The hash symbol (#) means "a number of".

Apart from the highly influencing PI MODS consisting of moderators dedicated particularly to that community, we observed strong impacts from other communities, such as PARENTING (3 times), SEX & SEXUALITY (18 times), or MODERATORS (6 times), but less frequently. The 6 strong impacts from MODERATORS were observed between weeks 151 and 246, as illustrated by Figure 12. From week 247 onward there were no further strong impact values identified which means that influence of MODERATORS on PERSONAL ISSUES lowered. However, that does not mean the forum stopped to be moderated, because a specifically-dedicated community of moderators PI MODS was established and had a strong impact in 24 cases from week 299 until the end of the data. The analysis of the time-series of the impact thus revealed which communities influenced PERSONAL ISSUES the most and when.

We also observed 10 strong impacts from SUPER MODERATORS on PI MODS, but in distinct weeks. Nevertheless, this suggests a presence of a hierarchy of influence: SUPER MODERATORS influenced another *moderating* community.

Influential Communities Are Often Private

Many of the influential communities were formed within private fora, i. e. fora with restricted access. We manually determined accessibility of all the communities for which we observed a strong impact relation in at least 10 weeks.

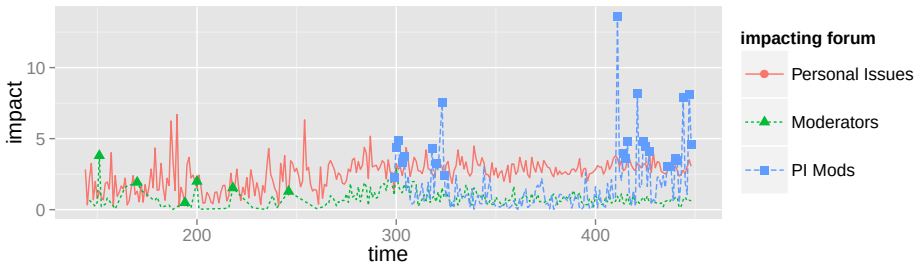


Figure 12: Impact of MODERATORS and PI MODS on PERSONAL ISSUES, and its independence (red solid line) over the time. The strong impacts from MODERATORS and PI MODS are emphasised by triangles and squares respectively.

The day of the analysis was 22. 5. 2013. Out of 25 influencing communities in total, 10 were public and 10 private, while out of 21 influenced communities 18 were public and 2 were private. Status of six communities could not be decided as they have been deleted. We removed PI MODS from the analysis, because it was both a subject and an object of a strong impact. We found out that the influential fora were significantly more likely to be private than the influenced communities (Fisher's exact test, $p = 0.004$). This suggests that in some cases the influential users group into small (compare the number of users in Table 3) elite communities that strongly influence dynamics in other communities.

While many of the strong impacts were induced by moderating or administrating fora, the impact is not restricted only to those cases, e.g. the impact of FNWAI on POKER. Even though the cross-community influence seems to be a more general phenomenon reflecting user grouping behaviour and interactions, we acknowledge that the results presented up to here are nevertheless based on our subjective interpretation. For that reason, we conducted a complementary analysis.

Strong Impact Indicates Higher Information Diffusion

In order to further validate our hypothesis that the strong cross-community impact indeed measures influence of one community on another, we utilised automated text analysis aiming to correlate language diffusion with the cross-community impact. Language and information diffusion has been commonly used as an indicator of social influence (see Section 2.5.2 on page 31) and it was found to be correlated with other aspects of leadership within online discussion communities, such as the ability to spark longer conversations or receiving replies [45].

Our hypothesis is that the information contributed by the members of influential communities diffuse more than the information contributed by the other members. In order to test that hypothesis, we adopted the *language diffusion score* $ld(i, v)$ of user i in community v proposed by Huffaker [45]:

$$ld(i, v) = \sum_{x \in P_{iv}} \sum_{y \in \Gamma(x)} \frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(x)|}, \tag{10}$$

where P_{iv} is the set of posts contributed by user i to community v; $\Gamma(\cdot)$ maps a post x to all the descendant posts in the conversation thread; and $\alpha(\cdot)$ maps a post to a set of keywords. In our experiments $\alpha(\cdot)$ was a function that removed any text quoted from a post x, tokenised the remaining text [78], filtered out any stop-words [73], and finally stemmed the remaining words [85]. The language diffusion score thus measures to what extent other members of community v adopt and replicate the terms used by user i. In other words, to what extent the information, measured as a number of keywords, contributed by user i cascades through the rest of community v. The divisor normalises the factors by the size of the keyword set of each post x, which corrects for a possible bias towards users who send long messages.

However, the score may remain biased by long replies, or by a high number of posts user i contributed to the community. We tried also normalising the numerator by the total number of keywords in both posts, $|\alpha(x) \cup \alpha(y)|$, which turns the fraction into a Jaccard index. We also tried to normalise it even further by the total number of posts user i contributed to forum v. Although the modified scores were generally lower that the original from Equation 10, they led us to the same conclusion: strong cross-community impact indicates higher information diffusion.

More concretely, we found that the members of community u that strongly impacts another community v have significantly higher average language diffusion score within v than the rest of the members of v (Wilcoxon signed rank test [104], $p < 2.2E^{-16}$). An experiment with median scores led to the same conclusion. This suggests that the strong impact indeed captures a tendency of a community to influence dynamics of another community, and in particular that the cross-community impact, as measured by COIN, may be a promising heuristic enabling efficient information diffusion. We explore this implication in Chapter 5.

INFORMATION DIFFUSION AND NESTED COMMUNITIES The language diffusion score as defined in Equation 10 may sometimes indicate the influence between communities spuriously if the communities are nested. Imagine for example a situation where all the members of smaller community u are also members of bigger community v, i. e. community u is *nested* in

community v: $C_u \subseteq C_v$. Let us further assume that the members of u frequently reply to each other within community v, but they do not interact with the other members of community v. Then a high language diffusion score $ld(i, v)$ of an overlapping member $i \in C_u \cap C_v$ in community v is induced by the replies only from the other overlapping members, i. e. $(C_u \cap C_v) \setminus \{i\}$. Therefore, it does not indicate an influence of community u on the non-overlapping part of community v, i. e. $C_v \setminus C_u$. In order to quantify to what extent community u influences only the non-overlapping part of v, we derive a *relative language diffusion score* that measures to what extent the keywords contributed by any actor i to community v cascade through the rest of the non-overlapping members, i. e. $C_v \setminus C_u$:

$$rld(i, u, v) = \sum_{x \in P_{iv}} \sum_{y \in \Gamma'(x)} \frac{|\alpha(x) \cap \alpha(y)|}{|\alpha(x)|}, \tag{11}$$

where $\Gamma'(x) = \Gamma(x) \setminus \cup_{j \in C_u} P_{jv}$ maps a post x to all the descendant posts in the conversation thread that were contributed by the non-overlapping members of u, i. e. $C_v \setminus C_u$.

We repeated the experiments using the relative language diffusion score, but although we observed the relative score values to be generally lower than the values of the original score (Equation 10), our findings were still statistically significant and consistent with the results for the original score. Even though the nested communities do occur in the data (e. g. PI Mods and Personal Issues), it seems that the nested structure does not confound our measurements. However, this may not hold in general and therefore we suggest to use the relative language diffusion score to avoid any potential issues with nested communities.

Actor-level Analysis May Be Misleading

One plausible and simple approach to cross-community influence analysis may be to first find the highly central individuals within the community and then to find their focal communities. However, such purely actor-level approach may lead to misleading conclusions, as we illustrate again on the example of Personal Issues and the moderating community PI Mods.

We compare the values of membership $\mathbf{M}_{iu}$, in-degree $\mathbf{C}_{iu}$, and language diffusion score $ld(i, u)$ of two groups of members of Personal Issues (community u) in week 448. The first group, denoted as *PIM*, consists of the users that are both members of Personal Issues and PI Mods, i.e. the overlapping users of the two fora. The second group, *PI*, consists of the rest of the members of Personal Issues. Figure 13 depicts the distribution of the values within the two groups as boxplots. We see that while the users from

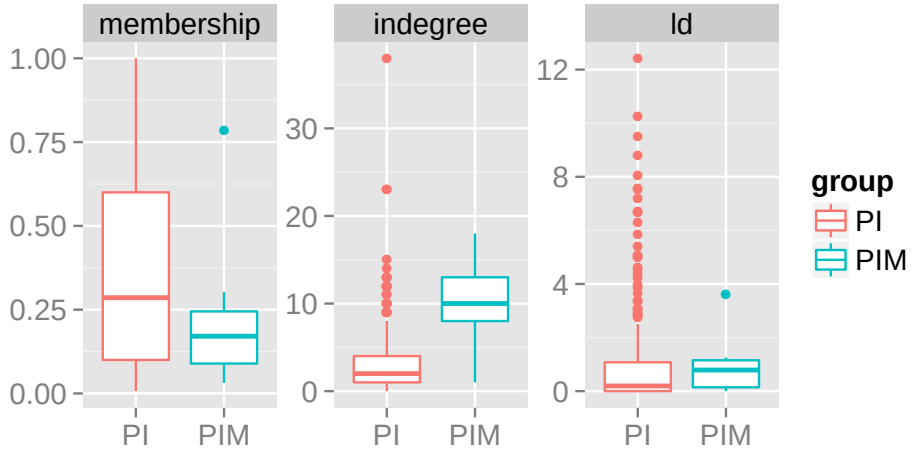


Figure 13: Values of membership, in-degree, and language diffusion score (*ld*) of the users from PI and PIM groups within the forum Personal Issues (see the main text).

the PIM group had much lower membership in Personal Issues, their median in-degree and language diffusion score were higher than those of the users from the PI group. In spite of that, some of the users from the PI group had much higher in-degree or language diffusion score than the users from PIM. We measured the membership of the users from the PI group with *high* in-degree and *high ld*, i.e. higher than a 1.5-times of the inter-quartile range, as denoted in the plot by the ends of the upper whiskers. The median membership for the users from PI with high in-degree was 0.38, with high *ld* it was 0.33, whereas the median membership of users from PIM was 0.17. This means that while PIM as a community had generally high influence on the rest of the Personal Issues, there were a few individual actors outside PIM who had very high in-degree or high *ld* in the community. A simple analysis focused only on the highly central actors and their membership independently may thus lead to deceptive results.

4.2.2 *Influence Between Pairs of SAP Communities*

We observed only a few strong impacts among the SAP communities, which is in accordance with the lower cross-posting activity, as we discussed in Section 4.1. Only for two pairs of communities we observed the impact to be strong in at least three different snapshots: an impact from Organizational Change Management to Business Process Expert General Discussion (4 strong impacts), and from SAP Business One—SAP Add-

ons to SAP Business One SDK (3 strong impacts). Forum Organizational Change Management is centred around topics related to organisational changes and business process re-engineering. Apart from its impact on Business Process Expert General Discussion we also observed a strong impact on another related community: Business Process Modeling Methodologies, although only in two snapshots. This suggests that the members of Organizational Change Management were important experts on business processes, because the impact for SAP is measured as the ability of one community to answer questions within another community. Similarly, the forum SAP Business One—SAP Add-ons, whose focus is development and deployment of extensions of the SAP system, had an impact on another forum with a similar topic: SAP Business One SDK (standard development kit).

4.2.3 *Summary of the Pair-Wise Analysis*

Analysis of strong cross-community impact on the level of pairs of communities revealed communities with high influence on other communities. In Boards, the influential communities were often private moderating or administrating fora, whereas in SAP we found only a few strong impacts between communities that both have similar focus. In general, we observed the impact values for SAP communities to be low in comparison with Boards communities. This corresponds to our previous observation of lower cross-posting activity in SAP (Section 4.1).

Although the pair-wise impact analysis delivers fine-grained insights into cross-community influence, it cannot easily reveal its overall trends. For example, in order to find the Boards community that was the most influenced by the other communities, i. e. the most *dependent* community, we may find the community that has received the most of the strong impacts and conjecture that it was After Hours. However, we defined aggregate measures specifically tailored for such analysis and the next section shows how we used them to track the overall trends of cross-community influence over the time.

4.3 OVERALL IMPORTANCE AND DEPENDENCE OVER TIME

Although the analysis of strong impact unveils cross-community influence relationship on the level of individual pairs of communities, it is hard to generalise the pair-wise analysis and answer *questions* like:

- Which communities had the highest impact on the rest of the communities?
- How did that change over time?

To answer these questions, we have to move our focus from the local level of pair-wise impact to its aggregates. Even the weak impact, if aggregated, may still provide interesting insight into global position of each community. For example, an otherwise independent community may have a role of a hub, which we expect to be indicated by a high total impact the community receives (see Proposition 2 on page 46).

Two aggregate measures were defined in Section 3.3 on page 42. The *importance* of a community (Equation 6 on page 44) is the total impact the community has on the other communities. The *dependence* of a community (Equation 7 on page 44) measures the total impact the other communities have on the community. Importance and dependence, along with their entropies (Definition 11 on page 45) measuring their heterogeneity, characterise overall cross-community impact for each community.

We investigate the importance and dependence values of Boards and SAP communities in three periods of length approximately one year: the *early* period corresponding to the beginning of the system, the *middle*, and the *late* period representing the most recent data we have.

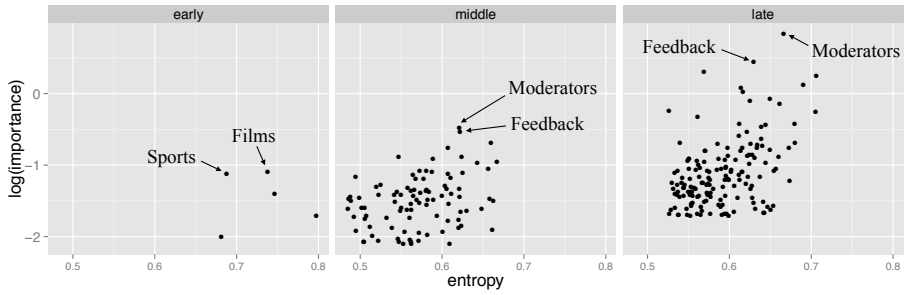


Figure 14: Importance and its entropy of all the fora in the early, middle, and late periods of Boards. For the sake of brevity, only the fora with at least median importance and its entropy are displayed.

4.3.1 *Importance and Dependence of Boards Communities*

IMPORTANCE Since the Boards data were segmented into weekly snapshots, each of the periods was 50 weeks long.[1] Figure 14 lists three plots, one for each period (column), depicting for each community its mean importance, logarithmically scaled for better comparison, along with the mean importance entropy. Please note that for the sake of brevity only the fora with at least median importance and entropy are displayed. The farther the community is from the origin, the more overall impact it had and the broader the impact was.

We see that in the early period the importance values were relatively low. In 19 out of 50 weeks of the period, the communities that had the maximal importance were centred around general topics: FILMS (10 times) and SPORTS (9 times). This means that in the early beginnings of Boards, when there were only a few communities with 699 users in total, the impact between communities was generally low, and thus the cross-community impact was less stratified.

However, this has started to change in the middle period, in which there were more users (8,069) and more communities, yet whose importance values did not grow substantially. There were no communities persistently having the highest importance. The community of MODERATORS, which had the highest importance most frequently in that period, had it only three times.

But that has changed in the late period. The number of active users grew to 36,474 and the importance of some of the communities rose notably (recall that the importance in the figure is logarithmically scaled) and became more stable: MODERATORS had the highest importance in 25 out of 50 weeks. The Figure 14 thus offers a condensed picture of the evolution of Boards: from the beginnings of the low cross-community impact heterogeneity and no dedicated administrating communities to a mature, large, and moderated system, in which there were communities with a broad range of cross-community influence.

DEPENDENCE Analysis of the community's dependence, reveals a different view on the Boards' evolution. Figure 15, analogously to the previous plot, depicts the mean dependence value and the mean dependence entropy for each community and period. The farther the community is from the origin, the more it was influenced by many other communities.

We see that in the early period the forum QUAKE had the highest mean dependence. The dependence values of QUAKE were the highest out of all

1 The early period spanned 12. 7. 1999–25. 6. 2000, the middle 5 .5. 2003–18. 4. 2004, and the late period was between 26. 2. 2007 and 10. 2. 2008.

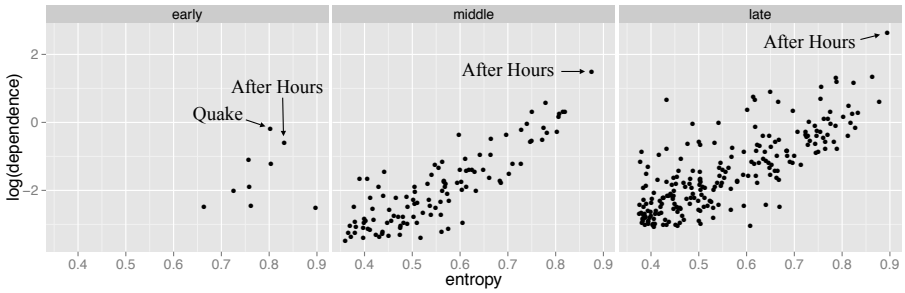


Figure 15: Dependence and its entropy of all the fora in the early, middle, and late periods of Boards. For the sake of brevity, only the fora with at least median dependence and its entropy are displayed.

the communities in 31 out of the 50 weeks of the period. This suggests that QUAKE, one of the first fora of the system altogether, served not only as a place to discuss the at that time popular computer game, but that its highly central users were participating a lot in other communities as well. Another highly dependent community was AFTER HOURS, that had the highest dependence in 13 weeks of the period.

This has started to change in the middle and especially in the late period, in which AFTER HOURS had persistently the highest dependence out of all the communities in 43 weeks of the middle period, and in 49 weeks of the late period. The mean dependence of AFTER HOURS also grew 25-times between the early and late periods, indicating that the conversational dynamics in the forum were increasingly being driven by users who belonged mainly to the other communities, and for whom AFTER HOURS served the purpose it was indeed founded for—a common meeting place.

Boards was founded initially as a discussion system for players of computer games and thus the high dependence of QUAKE in the early period indicates that it initially had a similar position in the system—a hub of the system. However, QUAKE had a stronger core of members in contrast with the over-arching AFTER HOURS. This can be measured as a ratio of dependence (Equation 7 on page 44) and independence (self-impact). The higher is the ratio, the more is the community dependent on the activity of focal members of other communities. The average ratio of dependence and independence for QUAKE was 2 in the early period, whereas for AFTER HOURS the same figure was more than 5-times higher in the middle and 12-times higher in the late period. AFTER HOURS thus emerged over time as a forum that is dependent on the activity of focal members of other communities. We observed a similar pattern in SAP as well.

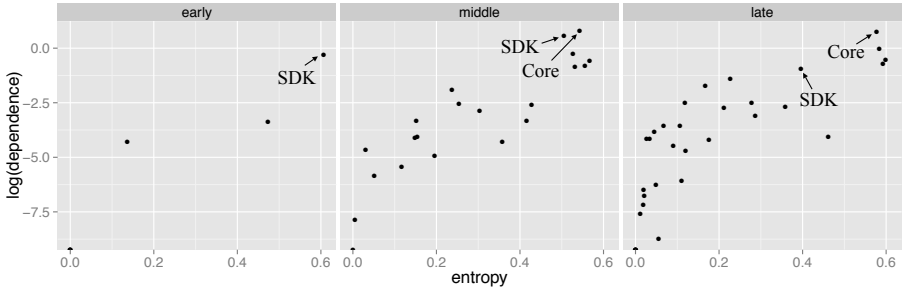


Figure 16: Dependence and its entropy of all the fora in the early, middle, and late periods of SAP. For the sake of brevity, we left out the leading "SAP Business One" from the titles of some of the fora.

4.3.2 *Importance and Dependence of SAP Communities*

In the previous Section 4.2 we saw that there were only a few strong impacts between the SAP communities. We now show that the analysis of aggregated measures, which take into account both strong and weak impacts, provide insight into the community dynamics on the global scale. Similarly to the analysis of Boards, we took data from early, middle, and late periods, each of which spanned six bi-monthly snapshots.[2] We found that the importance values were relatively low, not showing any rise of global authorities like in the case of Boards. This indeed is a consequence of only a few strong impacts between the communities and generally lower cross-posting activity (see Section 4.1), thus we omit the plots of importance. However, the analysis of dependence sheds more light onto which communities played the role of central hubs.

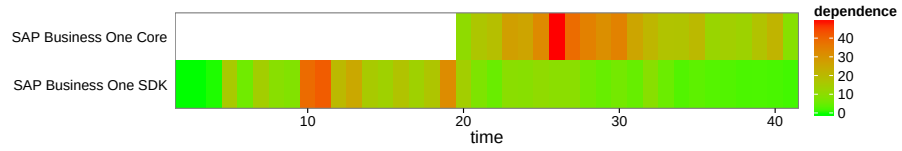


Figure 17: Dependence of the two SAP fora whose dependence was the highest in at least two snapshots.

2 The early period spans 1. 5. 2004–20. 4. 2005, the middle spans 1. 5. 2007 and 30. 4. 2008, and the late period is between 1. 3. 2010–28. 2. 2011.

DEPENDENCE The evolution of dependence depicted in Figure 16 shows that in the early period the most dependent forum was SAP BUSINESS ONE SDK (SDK) with the highest dependence values in 4 out of the 6 snapshots. However, in the middle and especially in the late period, the community SAP BUSINESS ONE CORE (CORE) became the most dependent.

This transition in dependence is also apparent in the heatmap in Figure 17. We see a sharp decline of dependence of SDK accompanied by the establishment of the forum CORE in snapshot 20, which has had a high dependence ever since. This suggests that the general discussion and questions related to the product SAP Business One moved at that time from the topic-specific SDK to the forum CORE, whose topic is not centred around any particular domain.

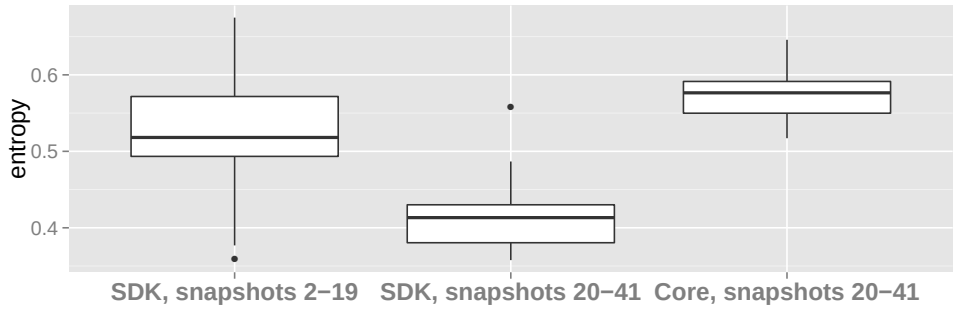


Figure 18: Dependence entropy of the SAP BUSINESS ONE CORE and SAP BUSINESS ONE SDK in the time snapshot 2–19 (only SDK) and 20–41 (both).

This is further indicated by the decline of the dependence entropy of SDK beginning with the time snapshot 20 as depicted in Figure 18. The figure compares the distribution of the entropy in two periods spanning the time snapshots 2–19 (SDK appeared first at time 2) and 20–41. We see that the entropy, and therefore the heterogeneity of the dependence, decreased for SDK, whereas it was notably higher for CORE. This means that the members of SDK became more focused, while the users of the new forum, CORE, participated also in a more diverse set of communities.

4.4 DISCUSSION OF THE RESULTS

In this chapter, we conducted a qualitative analysis of two different online discussion communities: Boards and SAP. While Boards is a general-purpose discussion system, SAP is technical support fora whose main purpose is to help its users to solve their technical problems. The flexibility of COIN allows us to take this difference into account and as a result we are able to identify

highly influential and highly influenced communities. The efficacy of COIN for quantification, analysis, and explanation of these phenomena motivated the integration of COIN into the community analytics platform PULSAR [70] developed by SAP.[3] Figure 19 shows a screenshot of the application.

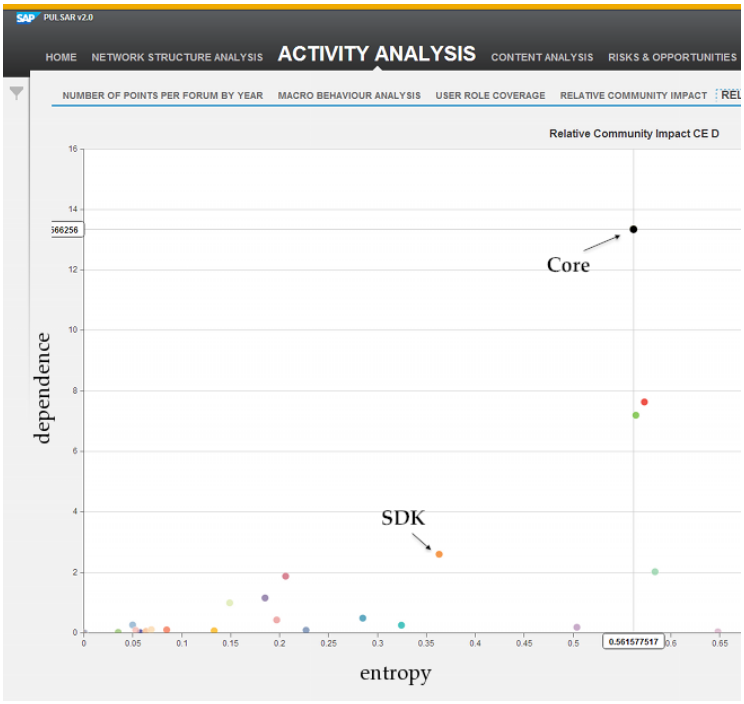


Figure 19: Screenshot of PULSAR application from SAP. The screenshot visualises the dependence (x-axis) and its entropy (y-axis) of the SAP communities in 2010. The plot in the screenshot is therefore similar to the subplot for the late period in Figure 16. However, unlike Figure 16, the plot in the screenshot has both axes linearly scaled.

We observed that Boards communities evolved from the beginnings featured by low activity, small number of communities, and no explicit moderating or administrating communities into a large, mature system, whose activity is highly influenced by multiple moderating, administrating, or expert communities, many of which were private. Some of the highly influential communities were relatively small, while the biggest and busy forum After Hours was found to be highly influenced by other communities. We found that similarly to the emergence of After Hours as the most dependent forum of Boards, SAP Business One Core appeared as the general purpose

3 PULSAR stands for Pulse Check Application for Online Communities.

community, which relies to a certain extent on the activity of the members of other communities.

Kraut and Resnik argued that high topic heterogeneity may contribute to a community's decline [53, p. 99]. A common way of reduction of that risk is to establish a community with a broad focus and to move into the community any topically heterogeneous, "off-topic", activity that may otherwise have a negative effect in more focused communities [53, p. 132]. The dependence and its entropy may thus serve as an early indicator of the lack of a community's focus and in turn it may help in decision making of the community's stakeholders. For example, they may decide to create a general purpose community like After Hours or SAP Business One Core in order to prevent the other communities from going off-topic.

Cross-community influence seems to be a general phenomenon reflecting user grouping behaviour and interactions. Whereas conventional measures cannot reveal this phenomenon and their usage may even lead to misleading conclusions, COIN enabled us to capture and analyse cross-community influence in two different online environments. Apart from exploratory and analytic purposes, the influence between communities may be leveraged for management of the communities, or it may be a useful feature for predictive analytics. For instance, a growth of a community's dependence may indicate an underlying shift of allegiance of its users, which may be considered undesirable by the community's stakeholders [50]. We also observed that strong cross-community impact indicates higher diffusion of language, which forms a basis for exploitation of the impact for efficient information diffusion. We explore this possibility in the next chapter, while we leave the other options for future work.

5

CROSS-COMMUNITY INFLUENCE AND INFORMATION DIFFUSION

We now move our focus beyond the exploration of the cross-community influence towards its exploitation for efficient information diffusion. In Section 2.5.2 on page 31 we discussed the fundamental problem of information diffusion maximisation as a selection of a small subset of *seed actors* in a social network who can efficiently disseminate some information or behaviour over the network. The main assumption is that the information or behaviour diffuses from the selected, or synonymously *targeted*, actors to their neighbours and further over a large part of the network.

However, we argued in Section 2.5.2 that in many cases the information is shared to the whole community and not to the individual actors. For example, in discussion fora a post is contributed to the whole community of the forum's users and not to its individual members. Therefore, we formally define, analyse, and propose a solution to the *cross-community information diffusion problem* formulated as a selection of a small subset of *seed communities* to target such that the information diffuses over the network as much as possible.

Similarly to the previous chapter, our approach is purely structural. Therefore, we assume that the stimulus is relevant for all the actors and the problem is to find a set of highly influential communities *in general*. In the next chapter, we extend COIN by content analysis in order to capture topics that may underpin the influence between communities.

We simulate the diffusion process by extensions of two commonly used models, that were adopted in order to start the process from seed communities instead of actors. The simulation is run over a social network derived from responses between actors. We use three different heuristics for selection of seed communities. We evaluate the efficacy of the heuristics by measuring the total number of users that were activated at the end of the diffusion process, i. e. the *user adoption*. In addition, we also measure the total number of communities that were activated at the end of the diffusion process, i. e. the *community adoption*.

We prove that the cross-community information diffusion problem is NP-hard, but we derive a greedy hill-climbing heuristic with an approximability guarantee of at least 63% of the optimal user adoption. As we found the greedy heuristic computationally expensive, we experimented also with

two other heuristics: a COIN-based strategy we call *impact focus* and *group in-degree* (Equation 3 on page 3), and compared all three heuristics with a random baseline. We refer to the heuristics and the random baseline commonly as the *targeting strategies*.

We first present two standard models that were previously proposed for simulating actor-level information diffusion in Section 5.1. We derive the social network that we used in our experiments in Section 5.2. After that, we extend the two models to enable study of information cascades over communities in Section 5.3. In Section 5.4 we define the problem of cross-community information diffusion formally, prove its complexity under the extended models, and present the greedy approach to its solution. We describe the other targeting strategies along with the experimental setup of this study in Section 5.5. Finally, in Section 5.6 we present the result which we further discuss in Section 5.7. Please note that similarly to the previous chapter, we use the terms "user" and "actor" interchangeably.

5.1 DIFFUSION STARTING FROM SEED ACTORS

As we discussed in Section 2.5.2 on page 31 several models of how information or an action diffuses over a social network have been proposed. The problem of maximising the diffusion of information or influence was first introduced by Kempe et al. [51], who proposed two generalisations of many previously defined models—*Independent Cascade Model* (ICM) and *Linear Threshold Model* (LTM). We used these two models because they have been a popular choice for simulating information and influence diffusion in social networks [51, 23, 24]. This also allowed us to follow the suggested experimental guidelines and parameter settings. Both ICM and LTM model a diffusion process over a social network $G = (V, E)$, starting from a set of initially activated seed actors $L \subseteq V$. The process unfolds iteratively from L to the rest of the actors over the weighted ties E until it converges. A weight $\mathbf{W}_{ij}$ expresses a propensity of actor j to adopt information from i.

LTM works *deterministically* in a pull mode, whereby a non-active actor j at iteration t is activated if $\sum_{i \in N_j} \mathbf{W}_{ij} \geq \theta_j$, where N_j is a set of active neighbours of actor j in the previous iteration $t-1$, and θ_j is a threshold expressing how many neighbours of actor j have to be active in order to activate her. That is, the decision of whether an actor becomes active or not depends *only* on the weighted sum of her active neighbours and the threshold.

ICM proceeds *stochastically* in a push mode over a network and the ties' weights represent probabilities of the information transmission between the actors. At each iteration t, each actor i that has been activated in the previous iteration $t-1$ has *exactly one* try to activate each of her non-active neigh-

bours j, and she succeeds with probability $\mathbf{W}_{ij}$. If any of the neighbours of actor j succeeds, j becomes active, but the neighbours have *no* further chances to pass their activation in the following iterations.

5.2 DERIVING THE SOCIAL NETWORK FOR INFORMATION DIFFUSION

As we discussed in Section 2.5.2 on page 31, the conversational behaviour of actors may be used as a proxy for information flow. Since a reply of actor j to actor i is an explicit engagement with the initial post, it indicates that actor j has ingested the information contributed by actor i and thus it indicates a flow of information in the opposite direction from i to j [105]. Therefore, we derive the social network we use for our evaluation from the replies in a way that is illustrated in Figure 20.

We used the Boards data that was presented in Section 4.1 on page 49 for the evaluation. Our assumption that the reply behaviour is a proxy for (reverse) information flow between the actors might not hold in a question-answering system like SAP. Therefore, we did not use SAP for the evaluation. On the contrary, the results of the language diffusion analysis described in Section 4.2.1 on page 52 showed that strong cross-community impact between Boards communities indicates higher diffusion of information. Hence we believe our modelling assumptions are plausible for that data. For the sake of computational tractability, we analysed only the last 31 weeks between 16. 7. 2007 and 10. 2. 2008. This approximates the last 7 months of our data-set and therefore it is the most recent and reasonably stable representation of the system we have.

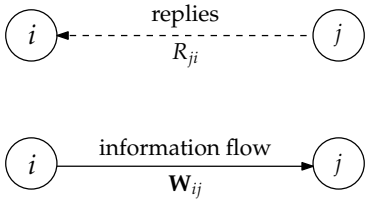


Figure 20: Illustration of how the information flows in the reverse direction of the replies.

We set the edge weights either globally to $z = 0.01$ and $z = 0.02$, because these values have been commonly investigated [23, 51], or we set them to the *likelihood* of the flow of information from actor i to actor j, $\mathbf{W}_{ij}$.[1] The

1 We also studied $z = 0.1$, but the diffusion process became insensitive to other parameters, such as the number of targeted communities, and reached the saturation point quickly. This was also observed by Chen, Wang, and Yang [23].

likelihood is calculated as the number of replies from j to i, R_{ji}, normalised by the total number of replies actor j posts:

$$\mathbf{W}_{ij} = \frac{R_{ji}}{\sum_{x=1}^{n} R_{jx}}, \tag{12}$$

where n is the total number of actors. We refer to the networks with global transmission probabilities $z = 0.01$ and $z = 0.02$ as $G_{0.01}$ and $G_{0.02}$. To the network with weights $\mathbf{W}_{ij}$ we refer as G_w.

5.3 DIFFUSION STARTING FROM SEED COMMUNITIES

In the LTM and ICM models described in Section 5.1 the diffusion process starts from the set of seed actors and *not* communities. We therefore extend the models such that the process starts from a set of *seed communities*, or simply *seeds* hereafter. Since communities consist of multiple actors and frequently share their members, i. e. they overlap, there are many ways to extend the models to the community level. For instance, if we stimulate a certain community, e.g. by posting into it, we could assume that *all* members of the community become activated and then the diffusion process unfolds, or that only a *subset* of the members becomes activated. If we assume that all the members are activated, it positively biases large communities over smaller ones regardless of the authority or participation patterns of their members. Moreover, this contrasts with the intuition that in a big community only a fraction of it would respond to the stimulus. This is because there is a higher likelihood that the stimulus would be missed by some members, e.g. those ones who only occasionally participate in the community. Therefore, from each of the seed communities we sample s *seed actors* and assume they responded to the stimulus, which then diffuses from them over the network.

5.3.1 *Sampling Seed Actors from Seed Communities*

Since we do not know how many actors would respond to the stimulus in the experiments, we investigated several sizes of the seed actors samples s in order to account for as many cases as possible. As communities may overlap, the samples from distinct targeted communities may overlap as well. This reflects the fact that the same actor may be stimulated in different communities. One way is to sample the seed actors uniformly, but this would not respect the distribution of responses of the actors. Indeed, actors who tend to respond more within the community are also more likely to respond to the stimulus. Hence we set the probability of actor i to be sampled as a seed

actor from community u to $\frac{B_{iu}}{\sum_{x=1}^{n} B_{xu}}$, where B_{iu} is the total number of replies user i contributed to community u, and n is the total number of users. If the number of members of the seed community was smaller than or equal to s, we took all its members.

5.3.2 *Extending ICM and LTM for Cross-Community Information Diffusion*

The extended *Community-Aware Linear Threshold (CALTM)* and *Community-Aware Independent Cascade (CAICM)* models proceed as follows:

1. Select set T of q targeted communities.
2. Obtain a final actor seed set L by sampling s members from each seed community u: $L = \cup_{u \in T} sample(u, s)$.
3. Run the original Independent Cascade or Linear Threshold Model (Section 5.1) with L as a set of seed actors.

The simulation was repeated h_d times for both models and we computed the expected value of user and community adoptions in order to account for the variance across the individual diffusion runs. In the case of CALTM, the reason is that we do not have the information about the thresholds θ in LTM that quantify the tendencies of actors to adopt the behaviour of their neighbours. Therefore, we set them uniformly at random for each run of the model, similarly to Kempe et al. [51], averaging over possible values of the thresholds. Likewise, ICM is a stochastic model and thus it is necessary to repeat the simulation to estimate the mean adoptions.

The community-aware diffusion models also involve sampling of seed actors from the seed communities. In order to account for the variance of user and community adoptions induced by the actor sampling, we estimate the adoptions for h_s independent user samples. Therefore, in total we run each diffusion model $h_s \cdot h_d$ times: h_d times (diffusion runs) for each of the h_s user samples.

5.3.3 *Measuring User and Community Adoptions*

Let A to be a set of all users that have been activated during the simulation. We measure the user adoption by *user activation fraction a* defined as $a = |A|/n$, the fraction of all the users that have been activated during the diffusion process. Analogously, the community adoption is measured by *community activation fraction c* defined as:

$$c = \frac{1}{k} \sum_{u=1}^{k} \left(\frac{\sum_{x \in C_u \cap A} \mathbf{M}_{xu}}{\sum_{x=1}^{n} \mathbf{M}_{xu}} \right), \tag{13}$$

where C_u is the set of all the members of community u and $\mathbf{M}$ is the membership matrix as defined in Section 3.3 on page 42. The numerator in the brackets sums up the memberships of the activated members of community u, and the divisor is the cardinality of the set representing the community. The fraction within the parenthesis thus quantifies the part of community u that has been activated. This is then summed over all communities and normalised by their total number k. The community activation fraction is 1 if all the users in all communities were activated, and 0 if no users were activated. Please note that c allows for the user activity patterns (by taking their memberships into account), and it also treats all the communities equally (by normalisation), which allows us to investigate diffusion across communities as opposed to individual users only.

5.4 MAXIMISING CROSS-COMMUNITY INFORMATION DIFFUSION

As the number of communities can be large, one cannot practically directly stimulate all of them. For instance, there is more than 600 communities in Boards (see Table 1 on page 51). Even if that was practically possible, such a strategy, close to aggressive spamming, would most likely be ignored. Direct addressing may also be costly. Ideally, the number of seed communities should be as low as possible, and therefore we define the problem as follows:

Definition 12 *The* cross-community information diffusion maximisation problem *is to find q communities, where q is minimal, s.t. the number of individual actors activated during the diffusion modelled by CAICM or CALTM is maximal.*

Since the problem is a generalisation of the previously studied influence maximisation by selecting seed *actors*, the lemma below follows from the previous theoretical analysis:

Lemma 1 *The maximisation problem from Definition 12 is NP-hard under both CAICM and CALTM.*

Proof 1 *Let the number of actors be the same as the number of communities, i. e. $n = k$, and let us assume that each actor belongs fully to exactly one community and each community has exactly one member. The problem then reduces to the influence maximisation by selecting seed actors, which is under both ICM and LTM a special instance of an NP-complete problem [51].*□

Despite the fact the problem is NP-hard, a simple greedy hill-climbing strategy offers approximability guarantee within at least 63% of the maximal user adoption. Kempe et al. [51] proved this for the original ICM and LTM

by analysis of submodular functions, i.e. a class of functions featured by the *diminishing returns* property. Concretely, they showed that the set function $\sigma(L)$, which maps the seed actors L to the expected number of actors that were activated at the end of the diffusion, is submodular. This means that for $L' \subseteq L$ the gain we get by including an additional actor into the seed set is gradually smaller as the seed set grows: $\sigma(L' \cup \{i\}) - \sigma(L') \geq \sigma(L \cup \{i\}) - \sigma(L)$. In the case of community-aware diffusion models, the set L is not fixed, but it is sampled from the seed communities. Therefore the expected number of all the actors that have been activated at the end of CAICM or CALTM is:

$$\sigma'(T) = \sum_{\text{all possible sampled users } L} p(X = L) \cdot \sigma(L)$$

where X is a random variable representing the seed actors sampled from the seed communities (see Section 5.3) and $p(X)$ is its probability distribution. Since the expected number $\sigma'(T)$ is a non-negative linear combination of submodular functions, it is submodular too [51]. Therefore a greedy hill-climbing strategy that iteratively selects the next best seed has an $(1 - 1/e - \epsilon)$-approximability lower bound on the found solution, where e is Euler's number and ϵ is an arbitrary positive precision parameter [74, 51].

GREEDY MAXIMISATION STRATEGY These results enable us to devise a simple seed selection strategy, which after q iterations returns a set T of q seed communities with the expected number of activated actors within at least 63% of the optimal number. In each iteration, the algorithm adds to set T community u that induces the maximal increment in the number of activated users: $\arg\max_u(\sigma'(T \cup \{u\}) - \sigma'(T))$. This corresponds to a greedy maximisation of the user activation fraction a, because a is the number of activated users $\sigma'(T)$ divided by constant n. In order to provide a fair basis for comparison with other heuristics, we do not maximise with respect to community activation fraction c, because that would yield two distinct sets of seeds: one maximising a and the other maximising c.

We have shown that the problem we study is a generalisation of the previously extensively studied problem of influence maximisation in social networks. Similarly to that problem it is NP-hard, but the greedy hill-climbing seed selection strategy promises to give satisfactory results. In spite of that, the greedy estimates can still be computationally expensive [23]. This is because in each iteration the greedy algorithm requires to run $h_s \cdot h_d$ diffusions in order to estimate the gain $\sigma'(T \cup \{u\}) - \sigma'(T)$ for each non-seed community u. Even though we reused some of the estimates [57], we still observed high running times in order of days.[2] Conversely, the other two heuristics,

2 For example, the greedy algorithm using CALTM model with G_w network run for 3 days on a machine with 12 Intel Xeon 2.4GHz cores.

impact focus and group in-degree, need to be computed only once for each community. Indeed, we observed running times in order of minutes for both of them. In the next section, we describe the other heuristics and the experiments we conducted in order to compare them with the greedy strategy.

5.5 EXPERIMENTS WITH TARGETING STRATEGIES

The main purpose of our experiments was to investigate the capability of the targeting strategies described below to maximise cross-community information diffusion with respect to three main factors:

1. Number of targeted communities (q)
2. Number of seed actors sampled from each targeted community (s)
3. Weights of edges in the information flow network

We evaluate the user and community adoptions induced by each of the targeting strategies by running a diffusion on either the same network snapshot t that was used by the strategy, or on the subsequent snapshot $t + 1$. We call the first scenario *seed selection*, while we refer to the other as *seed prediction*. Whereas the seed selection is the scenario commonly considered in the previous literature [23, 51], we believe that the seed prediction may be more appropriate in the real world. Indeed, most of the online communities are constantly changing, which may render any static representation of their underlying social structures quickly obsolete over time.

In total, we evaluated four *targeting strategies*:

1. *Greedy* (GR) finds seed communities by iteratively selecting the next best candidate community with the highest increment in the user activation fraction a (see Section 5.4).
2. *Impact focus* (IF) targets communities highly influencing many other communities. In order to find such communities, we propose to take a product of the importance (Equation 6 on page 44) and its entropy (Equation 8 on page 44). While the importance measures how much one community stimulates the other communities in total, the entropy captures how many distinct communities the community influences. Since a high importance of a community may be induced by strong impact relations with only a few distinct communities, we combine the importance with its entropy.
3. *Group in-degree* (GI) was considered as a reasonably well-established centrality measure of communities. It is defined as the number of ties

incoming to the members of the community from the non-members (see Equation 3 on page 18). Intuitively, it measures how much the community in total stimulates the rest of the system. We chose group in-degree because it is a generalisation of actor degree, which has been widely used as a heuristic for influence maximisation when targeting individual actors [51, 92]. Please note that the group in-degree, however, was not originally motivated by the influence maximisation problem and here it is used to represent an intuitive and simple heuristic only.

4. *Random* (R) was used as a baseline, and simply means a uniformly random choice of the communities to be targeted. For each combination of the number of targeted communities q and sampled seed actors s, we repeated the simulation for a different random sample of seed communities h_r times, and averaged the results. Random targeting, especially in combination with a high number of initially activated users, may be viewed as spamming.

The main parameters of the community-aware diffusion models are the number of targeted communities q and the number of users sampled from each targeted community s. We investigated up to five targeted communities, i.e. $q \in [1, 5]$, and $s \in [1, 10]$ seed actors sampled from each targeted community. We empirically observed the number of repetitions $h_s = h_d = h_r = 50$ led to the convergence of our measurements, while preserving computational tractability.[3] We conducted six experiments evaluating the suitability of each strategy to *select* seed communities, and another six experiments under seed *prediction* scenario. Each experiment corresponds to a combination of one of the three networks $G_{0.01}$, $G_{0.02}$, and G_w (see Section 5.3) and one of the two diffusion models. In each experiment, we investigated the user and community activation fractions induced by different values of q and s.

5.6 RESULTS

In this section, we report on the results of the experiments we conducted in order to analyse the information cascades across actors and communities. We used the two community-aware information diffusion models to simulate the cascades across the network and we measured the performance of each targeting strategy by the user activation fraction a and the community activation fraction c.

The main goal was to investigate which strategy performed the best in each of the experiments. In order to find only the cases in which a strategy

3 For example, the running time of all the experiments with the impact focus strategy took about 2 days and 15 hours using 4 Intel Xeon 2.27 GHz CPUs.

significantly outperformed the others, we employed statistical hypothesis testing. Specifically, in each experiment we identified two strategies that led to the highest user or community activation fractions: S_1 and S_2. S_1 was the strategy that led to the highest values of one of the performance measures for most of the combinations of the number of targeted communities q and the number of sampled seed actors s. S_2 was the next best strategy, i.e. the one which led to the highest values *excluding* the values achieved by S_1. The second strategy S_2 represents the upper bound of performance of the strategies that were alternative to the first strategy S_1. If S_1 led to significantly higher values of either of the performance measures a or c than strategy S_2, we conclude it was the best in the experiment with respect to the performance measure; otherwise we conclude that there was no such strategy. We test that hypothesis by a non-parametric Wilcoxon signed rank test [104]. The significance level was $\alpha = 0.05$ and family-wise error rate was controlled by Bonferroni correction. The test results are listed in Table 5. We discuss the results of the seed selection experiments first.

		CAICM			CALTM		
scenario	measure	G_w	$G_{0.01}$	$G_{0.02}$	G_w	$G_{0.01}$	$G_{0.02}$
seed selection	a	GR					
	c	GR	–	–	GR	–	–
seed prediction	a	IF	–	IF	IF	–	IF
	c	IF					

Table 5: The best strategy for each of the 6 experiments with seed selection and 6 experiments with seed prediction. No best strategy is indicated by a dash (–). If a strategy was the best in all the cases in one row, we list it only once.

5.6.1 *Results of the Experiments With Selection of Seed Communities*

In the six experiments under the seed selection scenario the greedy (GR) strategy was the best with respect to the user activation fraction a. GR was the best strategy with respect to the community activation fraction c too, but only for weighted networks G_w. Even though GR was the S_1 strategy also for networks $G_{0.01}$ and $G_{0.02}$, it was not significantly better than the S_2 strategy—impact focus (IF).

USER ACTIVATION FRACTION Since the results for the user activation fraction a were similar for all the seed selection experiments, Figure 21 illus-

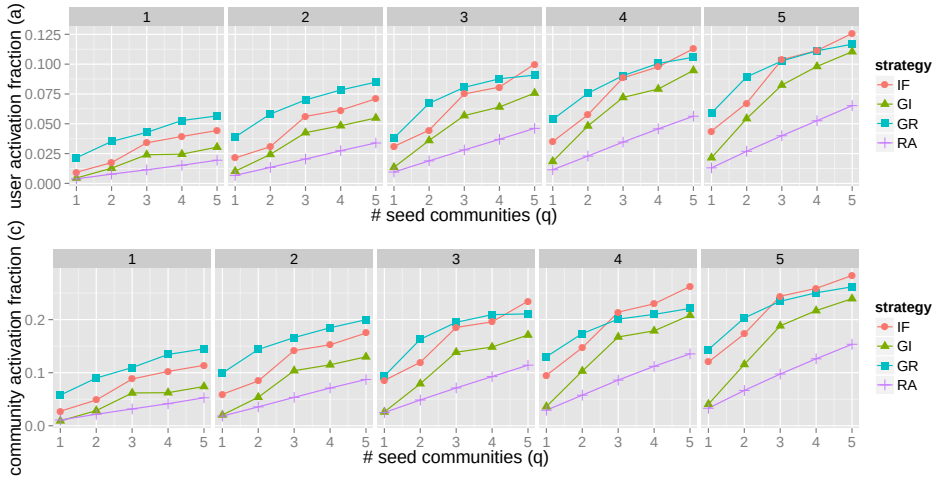


Figure 21: User and community activation fractions for the seed selection experiment with CAICM model and G_w from week 7. For the sake of brevity, only the plots for $s \in [1, 5]$ sampled seed actors are presented.

trates one representative example from the experiment with CAICM model on the weighted network G_w for the week 7 of the data (the complete set of plots for all the experiments is available online [12]). Each of the five subplots in the top row displays the mean user activation fraction (y-axis) induced by sampling s actors from each of the q targeted seed communities (x-axis) by each of the strategies: greedy (GR), impact focus (IF), group in-degree (GI), and random baseline (RA). For example, the leftmost subplot correspond to $s = 1$ actor sampled from each of the q seed communities. We see that the greedy strategy is the best, namely for small s and q. Please note the characteristic concave shape of the user activation fraction a for GR—a consequence of the submodularity of the estimated gains of a. Furthermore, we see that as the number of sampled actors s grows, the difference between the three heuristics diminishes—namely IF led gradually to similar a as GR. Although we present the results only for up to $s = 5$ for the sake of brevity, we observed similar trends in all our experiments up to $s = 10$.

COMMUNITY ACTIVATION FRACTION With respect to the community activation fraction c, GR was not significantly better than the S_2 strategy IF in experiments with globally set edge weights. The bottom row of Figure 21 depicts the mean community activation fraction c (y-axis), analogously to the upper row, with respect to the number of sampled actors s from each of the q seed communities (x-axis). We see that the greedy strategy led to the highest c for up to $s = 2$, but for a higher number of sampled seed ac-

tors, IF achieved higher c if at least three communities were targeted. We observed similar trends in the other experiments too and thus we believe it is the reason why GR was not significantly better than IF in these cases.

This on the one hand suggests that GR is especially suitable for situations when the likelihood that an actor responds to the stimulation is relatively low, i. e. low s, and the available resources allow targeting of only a small number of seed communities, i. e. low q. On the other hand, if the computational costs of GR prevent its usage, the impact focus offers a good alternative, especially if community adoption is the primary concern.

5.6.2 *Results of the Experiments with Prediction of Seed Communities*

GR was not the S_1 strategy in any of the experiments under the seed prediction scenario. IF was the best strategy with respect to both the user and community activation fractions, except the experiments on network $G_{0.01}$ (see Table 5). Despite IF was the S_1 strategy for network $G_{0.01}$, it was not significantly better than GI, the S_2 strategy.

We believe the cause of the worse performance of GR was overfitting. The *volatility* of the seed communities predicted by the greedy strategy was much higher than the volatility of communities predicted by IF or GI. In order to quantify the volatility, we computed the total number of *distinct* seed communities predicted by each heuristic over all the time snapshots, and compared its mean over each of the six experimental settings (choice of edge weights and diffusion model, see Section 5.5).[4] The volatility of GI was 45 and for IF it was 34, but for GR the volatility was more than 8-times higher: 287. That means that GR targeted a different community for almost every time snapshot and number of sampled actors s.[5] Conversely, GI and IF appeared to be biased towards targeting only a small number of specific communities, which made the two strategies more robust.

Similarly to the seed selection results, we generally observed consistent trends for all six experiments under the seed prediction scenario. Therefore we present only one representative Figure 22 that lists plots of the user and community activation fractions induced by s actors sampled from each of the q seed communities predicted for week 7 of the data (the plots for all the experiments are available online [12]). The 5 sub-plots in the top row correspond to $s \in [1, 5]$ actors sampled from the seed communities. We present only plots for s up to 5 for the sake of brevity, but the trends for higher s are similar. We see that especially for small s, GR yielded poor results, in

4 GI and IF target the same communities for all the values of s and diffusion models, whereas GR may target different seed communities for different diffusion models or their parameters.

5 Since seed communities were predicted for 30 time snapshots in total and $s \in [1, 10]$.

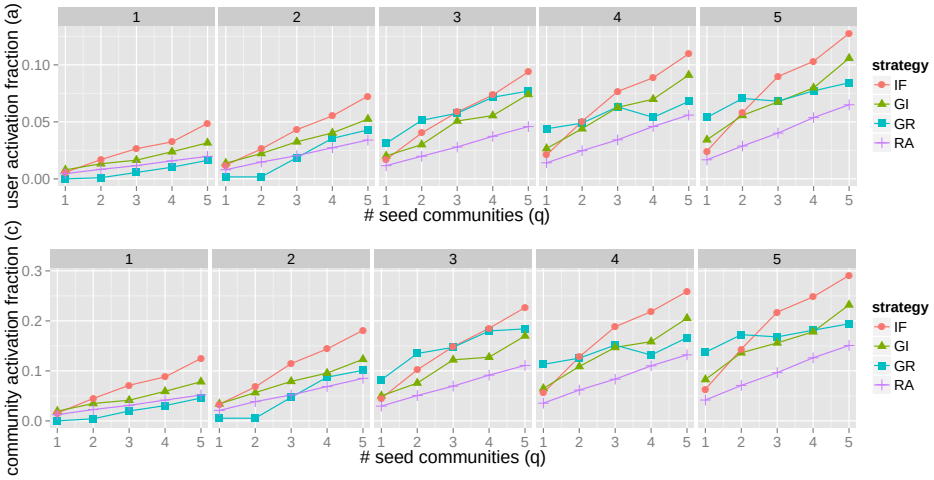


Figure 22: User and community activation fractions for seed prediction experiment with CAICM model on network G_w from week 7. For the sake of brevity, only the plots for $s \in [1, 5]$ sampled seed actors are presented.

some cases even worse than the random baseline. In contrast with that, GI and IF maintained user and community activation fractions similar to those achieved under the seed selection scenario.

5.7 CONCLUSION AND DISCUSSION OF THE RESULTS

We have motivated and defined the problem of cross-community information diffusion maximisation as a selection of a small set of seed communities to target such that the information diffuses over as many actors as possible. In addition, we also measured the total number of communities that have been activated during the diffusion process. We defined two community-aware diffusion models as extensions of two popular actor-level models: ICM and LTM. We proved that the problem is NP-hard under the extended models. We proposed two heuristics for its solution: a greedy hill-climbing strategy and a COIN-derived strategy that we call impact focus.

The results indicate that the greedy strategy is suitable for identifying seed communities in the presence of a relatively stable social network, especially whenever the likelihood of the actors to respond to the stimulus is low and the resource restrictions require targeting only as little communities as possible. However, this comes at the cost of the higher computational complexity and overfitting. In particular, we observed that the greedy strategy yields poor results if the underlying information flow network changes. Since many online communities are highly dynamic with their users join-

ing,leaving, and changing their mutual bonds, the greedy strategy seems not to be broadly applicable for such environments.

On the contrary, impact focus turned out to maintain persistently good results under both seed selection and seed prediction scenarios, and with respect to both user and community adoptions. Together with the fact that it is less expensive to compute compared to the greedy strategy, it appears to be a promising heuristic to consider whenever the conditions resemble those assumed in our experiments.

There are several ways how to build upon or improve the results presented in this chapter. The impact focus strategy may be improved by penalising overlap between the seed communities—a successful strategy proposed for the actor-level diffusion maximisation problem [23]. This may in effect lead to a broader coverage of different parts of the network and thus it may lead to higher user or community activation fractions. Furthermore, even though we observed in Section 4.2.1 that the information from the members of the influential communities diffuse more than the information contributed by the other actors, more research is needed to validate the community-aware diffusion models on real information cascades. In the more common case of the actor-level diffusion, Saito et al. [87] proposed a model selection approach that estimates the parameters of diffusion models from empirical cascades data. A similar extension for the cross-community cascades is thus one direction for future research.

Finally, we assumed that the stimulus is relevant for the whole system and therefore the problem was to find a set of highly influential communities in general. In the next chapter, we extend COIN by analysis of content contributed by the actors in order to investigate what topics are associated with an observed cross-community impact. This allows us to find highly influential communities with respect to a particular topic like "computer games".

6

TOPICAL DIMENSIONS OF CROSS-COMMUNITY INFLUENCE

The cross-community influence framework we have developed and evaluated in the previous chapters measures the impact in a purely *structural* form—without considering any additional information like topics of the discussions. We have demonstrated already that with the purely *structural impact* defined by COIN interesting insights into cross-community dynamics can be obtained. However, the interpretability of the structural impact is limited to domain knowledge and information like names of the communities, their size, or activity.

Sometimes additional information like textual data contributed by the actors may be available. For example, the messages contributed by the users of discussion fora represent a wealth of additional information about their interactions. While an actor may stimulate high activity, i. e. be influential, with respect to a particular topic like "computer games", she may receive few responses about another topic like "music". This also means that a community may be influential with respect to one topic, but not influential with respect to other topics.

We extend and generalise COIN in order to integrate information about topics extracted from the documents that were contributed by the actors, e. g. posts or messages in discussion communities. We refer to the original framework as *structural* COIN and we call the generalised framework *topic-informed* COIN. Likewise, we differentiate between *structural* and *topic-informed* impacts.

Using the topic-informed COIN, we investigate topics that may underpin the cross-community influence between pairs of Boards communities by analysing *topic-informed impact*. Similarly to our methods in Chapter 4, we evaluate the framework by qualitative analysis. We expect that the topic-informed impact will:

- provide better *interpretability*;
- be more *sensitive* if the influence relation between two communities is induced by activity associated with a particular conversational theme.

The remainder of the chapter is organised as follows. In Section 6.1 we introduce a few concepts from tensor algebra, which are required for the representation of the additional topical dimensions. After that, in Section 6.2,

we define the notion of *topic-informed impact* and present how it can be practically measured using a topic model. In Section 6.3 we demonstrate the efficacy of the topic-informed COIN on a follow-up analysis of Boards data that was presented already in Section 4.1 on page 49. We close the chapter with discussion of the limitations and possibilities posed by the generalised framework in Section 6.4.

6.1 SELECTED CONCEPTS FROM TENSOR ALGEBRA

The structural COIN represents the dynamics in the system using two dimensions only—the actors and their communities. It represents their memberships and centralities in the communities as matrices. In order to represent the system's dynamics with additional dimensions like topics, we need to move from the matrix-based representation to a tensor-based one.

A tensor is a generalisation of a concept of a vector, which is a mode-1 tensor, or a matrix (a mode-2 tensor) to an arbitrary m modes (dimensions). For example, while a vector can be imagined as a sequence of numbers and a matrix as a table, a mode-3 tensor can be conceived as a cube as depicted by Figure 23. For the sake of brevity, we denote vectors using small-case bold latin characters, e. g. $\mathbf{x}$, matrices as upper-case bold latin character, e. g. $\mathbf{X}$, and for tensors we use upper-case calligraphic latin characters, e. g. $\mathcal{X}$. In the same vein, an i-th element of vector $\mathbf{x}$ is $\mathbf{x}_i$, and an element in the i-th row and u-th column of matrix $\mathbf{X}$ is $\mathbf{X}_{iu}$. We will use the same sub-script notation for tensors of higher modes by simply adding more indices. In matrix algebra, we may obtain a vector of numbers by taking all elements of matrix $\mathbf{X}$ along one mode while keeping the other mode fixed: the vector $\mathbf{X}_{i\cdot}$ is commonly called row i and the vector $\mathbf{X}_{\cdot u}$ is called column u. Analogously, a *fibre* is a vector resulting from taking all elements of tensor $\mathcal{X}$ along one mode while keeping all the other modes fixed [9]. Figure 23 illustrates the difference between a matrix column and a column fibre of a tensor.

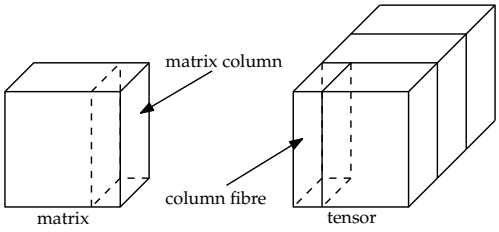


Figure 23: An illustration of a matrix column and a column fibre in a mode-3 tensor.

CONTRACTED TENSOR PRODUCT The only operation we need for our generalised definitions is the contracted tensor product. The general definition [9] is very flexible, but for the sake of clarity we provide only its simplified version here. In the following, we use capital letters with subscripts like I_1 to denote the size of the first mode of a tensor and small case letters like i_1 to denote the index variable along the first mode. Therefore, the indices along the first mode run within $i_1 \in [1, I_1]$. As the contracted tensor product that we introduce below contracts the first m modes of two tensors, it is useful to distinguish between the first m modes of a tensor and the rest. Therefore, we define the product for two general tensors $\mathcal{X}$ and $\mathcal{Y}$, where $\mathcal{X}$ is of size $I_1 \times \cdots \times I_m \times J_1 \times \cdots \times J_n$ and $\mathcal{Y}$ is of size $I_1 \times \cdots \times I_m \times K_1 \times \cdots \times K_p$.

Definition 13 *Let $\mathcal{X}$ to be a $I_1 \times \cdots \times I_m \times J_1 \times \cdots \times J_n$ tensor, and $\mathcal{Y}$ to be a tensor of size $I_1 \times \cdots \times I_m \times K_1 \times \cdots \times K_p$. By contracted tensor product we mean an operation denoted as $\otimes_m$ that multiples the two tensors along the first m modes. The resulting tensor $\mathcal{Z} = \mathcal{X} \otimes_m \mathcal{Y}$ is of size $J_1 \times \cdots \times J_n \times K_1 \times \cdots \times K_p$ [9]:*

$$\mathcal{Z}_{j_1 \ldots j_n k_1 \ldots k_p} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_m=1}^{I_m} \mathcal{X}_{i_1 \ldots i_m j_1 \ldots j_n} \mathcal{Y}_{i_1 \ldots i_m k_1 \ldots k_p} \tag{14}$$

Please note that from Definition 13 it follows for two matrices $\mathbf{M} : n \times k$ and $\mathbf{C} : n \times k$ that $\mathbf{M} \otimes_1 \mathbf{C} = \mathbf{M}^{\mathrm{T}}\mathbf{C}$. Therefore the structural impact matrix $\mathbf{J} = \mathbf{S}^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{C}$ (see Definition 6 on page 43) can be expressed using the contracted tensor product as $\mathbf{S}^{-1} \otimes_1 \mathbf{M} \otimes_1 \mathbf{C}$.[1] Similarly to our matrix-based definitions of impact, we will use the tensor product to combine information about the topics, memberships, and centralities of actors into a topic-informed impact tensor.

6.2 MEASURING TOPIC-INFORMED IMPACT

In order to measure an impact one community has on another community with respect to a particular topic, we extend the definitions of structural COIN from Section 3.3 on page 42. In Section 6.2.1, we generalise the definition of structural impact leveraging the concepts of tensor algebra introduced above. The structural impact is obtained by combining distributions of actors' memberships and centralities. In the purely structural approach the memberships and centralities are measured by the distributions of the actors' documents and the responses between them. In addition to that, the *topic-informed* COIN assumes that each document can be represented as a distribution over d topics. This means that each document is "divided" non-uniformly into d topics according to its content. For example, a message

Section 3.1 on page 37 explains how we derive the social network from the responses between documents.

1 Recall that $\mathbf{S}^{-1}$ is diagonal and thus $\mathbf{S}^{-1} = (\mathbf{S}^{-1})^{\mathrm{T}}$.

posted to a discussion forum about "sports" is likely to be mostly about "sports" and less likely about "music".

With this assumption, we can measure to what extent the author of a particular document stimulated responses about a particular topic like "music", i. e. her *topic-informed centrality*. Likewise, we can define the *topic-informed membership* by a level of activity associated with the topic in each community. We obtain the *topic-informed impact* as a combination of the distributions of topic-informed memberships and centralities.

In Section 6.2.2 we discuss our choice and calibration of a topic model and how we mapped the extracted topics to the notions of topic-informed membership and centrality. In the following, we consider a general case of n actors (users), k communities (fora), and d topics.

6.2.1 *Adapting COIN to Measure Topic-Informed Impact*

TOPIC-INFORMED MEMBERSHIP We argued earlier in Chapters 2 and 3 that an actor may be a member of multiple communities with a varying degree of membership. We measured the membership of an actor in a community as a fraction of all the actor's posts that the actor contributed specifically to the community. A high membership in one community expressed the actor's preference to contribute to that community rather than to the other communities. While this notion of an actor's membership provides a global picture of her preferences, there may be situations where it does not represent the preferences accurately. For example, an actor may prefer forum about "sports" overall, but to discuss "music", she is more likely to prefer a forum whose topic is closer to "music". In order to reflect this, we define a membership of actor i in community u with respect to topic e as a fraction of all her posts that she contributed to community u about topic e. Formally, the actors' memberships are represented as a $n \times k \times d$ **membership tensor** $\mathcal{M} : \mathcal{M}_{iue}$, such that $\forall_i \sum_{x=1}^{k} \sum_{y=1}^{d} \mathcal{M}_{ixy} = 1$. This condition requires that the membership of an actor is normalised over all the communities and topics. Each actor is therefore "divided" across multiple communities and topics.

The notion of topic-informed membership is a generalisation of the purely structural membership because if all the actor's activity is about one topic, the two types of membership are equal. Similarly to the purely structural membership, each fibre $\mathcal{M}_{\cdot ue}$ represents the fuzzy *set of members* of community u with respect to topic e. The sum over the fibre $\sum_{x=1}^{n} \mathcal{M}_{xue}$ represents the *cardinality* of the set. Analogously to the Proposition 1 on page 40, we propose to differentiate *focal* and *alter members* of a community with respect to their topic-informed membership:

Proposition 3 *Let u and v to be overlapping fuzzy communities according to Definition 3 on page 14. The topic-informed membership of each actor in each community is computed as described above. For any actor i whose membership is higher in community u than in v, i. e. $\mathcal{M}_{iue} > \mathcal{M}_{ivf}$, we propose to call:*

- *actor i a* focal member *of community u with respect to topic e and* alter member *of community v with respect to topic f;*
- *community u the* focal community *of actor i with respect to topic e;*
- *community v an* alter community *of actor i with respect to topic f.*

We define the set of actors for whom u is the focal community as the focal members *of community u with respect to topic e. Conversely, we define the set of actors for whom v is an alter community as the* alter members *of community v with respect to topic f.*

The notions of focal and alter memberships express the *relative* preference of an actor to contribute to the topics of two communities. In the simplest case, the two topics are equal, i. e. $e = f$, and thus the focal membership of actor i in u means that the actor prefers community u over community v to contribute to the topic. Furthermore, in Section 6.3.3 we show that investigation of differences between memberships informed by two distinct topics leads to explanatory insights into cross-community influence.

TOPIC-INFORMED CENTRALITY In addition to the notion of membership, the second ingredient to the measurement of cross-community impact is the distribution of actor's centrality in each community. In the structural COIN, an actor was highly central if she was able to stimulate high volume of activity within that community. We can again imagine situations, where the structural notion of centrality may not represent the dynamics as desired. An extreme example are the users of online communities (sometimes called "trolls" [53, p. 135]) who regularly contribute highly provocative messages that sway the dynamics towards different, often off-topic, conversations. Despite the fact that such users would have high centrality in the community, the centrality does not represent an ability to stimulate *desirable* activity in the community. In order to overcome this limit, we measure the centrality of actor i in community u with respect to each topic e. Formally, we represent the centralities as a $n \times k \times d$ **centrality tensor** $\mathcal{C} : \mathcal{C}_{iue}$. Similarly to our previous structural analysis in Chapter 4, we defined the centrality as an actor's in-degree (Equation 1 on page 18) with respect to each topic. An actor is therefore again considered as influential with respect to a topic if she *stimulates* many replies, i. e. high activity, about the topic.

Figure 24 illustrates the way we measure the topic-informed centrality. For the sake of simplicity, we assume only two topics e and f denoted as white (e) and black (f). There are two documents entirely about white topic e and three documents entirely about black topic f. However, in practise we expect the documents to have mixed topics. Since actor i stimulated one response with respect to each topic, her centralities $\mathcal{C}_{iue}$ and $\mathcal{C}_{iuf}$ are both equal to 1. Analogously, as actor j stimulated one response about topic f but no response about topic e, her centrality $\mathcal{C}_{iuf}$ is 1 but $\mathcal{C}_{iue} = 0$. Please note that since our definition of centrality informed by a topic measures only the ability of an actor to stimulate *responses* about the topic, we do not take into account the topic of the original (responded) document. The topics of the responded documents may be analysed by using a topic-relational model. We return to this subject in Section 6.4.

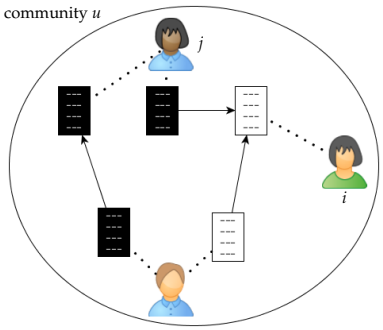


Figure 24: An example of the way we measure the topic-informed in-degree. A dotted line expresses an authorship of a document and a directed link represents a response between two documents.

TOPIC-INFORMED IMPACT Since both membership and centrality are now quantified for each topic, we can define the topic-informed impact $\mathcal{J}_{uevf}$ as a measure of the tendency of the focal members of community u with respect to topic e to stimulate activity about topic f in community v. In the simplest case, we can set the two topics equal, i. e. $e = f$. For example, if we set the two topics to "music" then a high impact of community u on v indicates that the actors who stimulate high volume of activity about "music" in community v are the focal members of u with respect to "music", i. e. they prefer community u over v to discuss "music".

Generally speaking, we can investigate impact between communities as induced by an interplay of different conversational topics. The caveat is that for d topics there are $d(d-1)$ possible topical interactions, which may quickly lead to practical infeasibility of the analysis. Moreover, not all combi-

nations are meaningful since we may expect that the activity in a community is related to only one or a few major topics. Instead of analysing every possible combination, our proposal is that the space of possibilities is pruned by expressing which combination of topics may feature interesting influence dynamics based on domain knowledge. Additionally, in Section 6.3.1 we demonstrate how to prune the space semi-automatically by selecting only the main topic for each community.

With this in mind, we are now ready to formalise the topic-informed cross-community **impact tensor** as a contracted product (Definition 13 on page 83) of the membership and centrality tensors. Analogously to the definition of structural impact, we normalise the topic-informed impact by the size of the impacting community:

Structural impact is defined in Definition 6 on page 43.

Definition 14 *We define the $k \times d \times k \times d$* impact tensor *as a contracted product of centrality and membership tensors:*

$$\mathcal{J} = \mathcal{S}' \otimes_2 \mathcal{M} \otimes_1 \mathcal{C} \tag{15}$$

where $\mathcal{S}'$ is a $k \times d \times k \times d$ tensor of reciprocal values of the community sizes with respect to each topic:

$$\mathcal{S}'_{uevf} = \begin{cases} \frac{1}{\max(\sum_{x=1}^{n} \mathcal{M}_{xue}, 1)} & \text{if } u = v \text{ and } e = f \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

Element-wise, the impact $\mathcal{J}_{uevf}$ of community u with respect to topic e on community v with respect to topic f is the normalised sum of centralities of the members of u within community v, weighted by their membership in u:

$$\mathcal{J}_{uevf} = \frac{\sum_{x=1}^{n} \mathcal{M}_{xue} \mathcal{C}_{xvf}}{\max(\sum_{x=1}^{n} \mathcal{M}_{xue}, 1)} \tag{17}$$

As follows from the definition of the contracted tensor product, all the tensor-based definitions are generalisations of the matrix notation introduced in Chapter 3. First, we compute the product of $\mathcal{M}$ and $\mathcal{C}$ along the first mode, multiplying the memberships of actors in community u with respect to topic e with their centralities in community v with respect to topic f. The result is a $k \times d \times k \times d$ tensor of unnormalised impacts—let us call it $\bar{\mathcal{J}}$. We further normalise $\bar{\mathcal{J}}$ by multiplying it with $\mathcal{S}'$. From the definitions of $\mathcal{S}'$ and the contracted tensor product it follows that $\mathcal{J}_{uevf} = \sum_{x=1}^{k} \sum_{y=1}^{d} \mathcal{S}'_{xyue} \bar{\mathcal{J}}_{xyvf} = \mathcal{S}'_{ueue} \bar{\mathcal{J}}_{uevf}$, because $\mathcal{S}'_{xyue}$ is non-zero only for $x = u$ and $y = e$ according to Equation 16. As a result, each impact value in the unnormalised impact tensor $\bar{\mathcal{J}}_{uevf}$ is divided by the size of the community u with respect to topic e. As we already discussed in Section 3.3 on page 42, a cardinality of the fuzzy

set representing the impacting community u can be lower than 1. In such cases, by dividing the unnormalised impact by a size that is lower than 1 the impact would be inflated. In order to address this issue, we again introduce a community size threshold $\theta = 1$ below which we do not normalise the impact.

INDEPENDENCE AND STRONG IMPACT Analogously to our earlier definitions from Section 3.3 on page 42, we call the self-impact $\mathcal{J}_{vfvf}$ of community v its **independence** with respect to topic f. The independence expresses to what extent the focal members of community v generate activity about topic f in the community. Analogously to Definition 10 on page 44, we further differentiate between *strong* and *weak* topic-informed cross-community impact:

Definition 15 *We call the impact $\mathcal{J}_{uevf}$ of community u on community v* strong *if it is at least as high as the independence (self-impact) of the community v, i.e. if $\mathcal{J}_{uevf} \geq \mathcal{J}_{vfvf}$. Otherwise we say that the impact is* weak.

6.2.2 *Extracting Topics for Cross-Community Influence Analysis*

Our definition of topic-informed impact is very flexible and promises to be compatible with many of the currently available topic models [44, 16, 28, 29, 22]. However, the discussion of the advantages and disadvantages of different topic models is out of the scope of this work. In this chapter, we use Latent Dirichlet Allocation (LDA) for the evaluation, but we believe that interesting insights may be obtained by applying the other models as well. We chose LDA because it is a well understood and widely used topic model that is implemented in freely available and scalable software [67].

LATENT DIRICHLET ALLOCATION Latent Dirichlet Allocation [16] is probably one of the most popular topic models at present. Given a set of documents P (corpus) and a number of topics d, it characterises the documents as a mixture of two probability distributions. The first distribution represents the probability of a word given each topic. The second distribution represents a probability of a topic given each document. Formally, we denote the second distribution as $p(Z|D)$, where D and Z are random variables representing the documents (D) and the topics (Z). Therefore, the higher is the probability of topic e given document l, $p(Z = e|D = l)$, the more likely are the words characterising the topic e to occur in the document.

TOPIC MODEL CALIBRATION We fit LDA to all the posts contributed by the users of Boards using machine learning package Mallet [67]. Each post

was represented as a separate document. In order to improve the quality of the topics, we merged the title and the body of each post and converted the text into lower-case; tokenised it [78]; we removed any stop-words [73] or words with less then 3 occurrences in the corpus; and finally we stemmed each token [85]. By a manual inspection of the titles of the fora we estimated the number of the high-level topics to be between 20 and 30. We therefore experimented with $d \in \{20, 30, 40\}$ topics, but since the results were similar and already satisfactory for $d = 20$, we used the model with 20 topics in our evaluation. Table 6 lists the topics along with the 7 words with the highest probability given each topic. Please note that due to the stemming, some tokens are no longer valid English words, e. g. "plai" (original "play") or "peopl" ("people"). In contrast with the rest of the topics, we found the topic "general" noisy and not very informative as it consists of many general or unrelated words. Since the "general" topic is associated with 21% of all the activity, we deem it as an unrepresentative noise and we omitted it from any further analysis.

TOPIC-INFORMED MEMBERSHIP Since LDA divides each post across d topics by the distribution $p(Z|D)$, we can use the distribution to measure the users' memberships with respect to each topic:

$$\mathcal{M}_{iue} = \frac{\sum_{x \in P_{iu}} p(Z = e|D = x)}{\sum_{x=1}^{k} |P_{ix}|}, \tag{18}$$

where P_{iu} is the set of all the posts that user i contributed to community u. We again obtain the purely structural form as a special case if all the posts P_{iu} of user i in community u are entirely about topic e.

TOPIC-INFORMED CENTRALITY We use the distribution $p(Z|D)$ also to measure the centralities of users within the communities. More concretely, as the centrality $\mathcal{C}_{iue}$ of user i should measure the tendency of the user to *stimulate* conversation about topic e within community u, we define the centrality as $\mathcal{C}_{iue} = \sum_{x \in F_{iu}} p(Z = e|D = x)$, where F_{iu} is the set of all the replies to the posts that user i contributed to community u. This is again a generalisation of the structural model from Chapter 3 in the sense that if all the replies within community u were entirely about one topic e, then the topic-informed and structural centralities are equal.

topic	prob.	words
religion	0.02	god,christian,religion,human,peopl,church,exist
gambling	0.02	call,hand,plai,fold,bet,player,rais
Ireland	0.03	peopl,irish,countri,ireland,war,world,american
motoring	0.04	car,drive,road,driver,speed,insur,engin
games	0.04	game,plai,server,good,player,xbox,map
politics	0.04	law,govern,compani,public,servic,ireland,year
music	0.04	music,band,plai,song,album,gig,sound
school & work	0.04	year,work,student,school,colleg,job,train
sports	0.04	team,plai,game,player,win,season,year
electronics	0.04	card,sky,channel,box,cabl,nokia,drive
food	0.05	eat,water,good,weight,food,drink,dai
shooting	0.05	dog,back,gun,shoot,shot,time,fire
cities	0.05	dublin,citi,road,area,place,bu,street
films	0.06	film,show,watch,movi,good,episod,book
internet	0.06	connect,work,file,problem,set,instal,download
money	0.08	price,pai,bui,month,monei,phone,offer
moderating	0.08	post,thread,forum,board,ban,peopl,http
relationships	0.10	peopl,girl,thing,friend,gui,time,make
argumentation	0.10	peopl,make,point,thing,time,good,work
general	0.21	good,time,dai,work,back,dont,thing

Table 6: List of the topics extracted from the Boards data. The topics are presented in the ascending order of their probability in the whole corpus, i. e. the most specific topics are on the top. We labelled the topics manually by inspection of the full list [12] of the top words for each topic as provided by Mallet [67].

6.3 ANALYSIS OF TOPIC-INFORMED IMPACT BETWEEN BOARDS COMMUNITIES

We applied the topic-informed COIN on Boards data that was already presented and analysed using the structural COIN in Chapter 4. Our aim was to demonstrate that the topic-informed COIN can reveal new insights that explain the cross-community influence.

Each Boards forum is typically centred around one or a few specific topics. Even though we observed that LDA assigned most of the probability mass for each post in a forum to typically one or only a few of the topics, it

frequently assigned a non-zero probability to many of the topics. While the high-probability topics typically represent meaningful information, the low-probability topics of a post can be considered to be not much more than a statistical noise. Therefore, in order to focus our analysis only to meaningful relationships between the communities, we first determine the main topics for each community in Section 6.3.1.

We approach our analysis of topic-informed impact between Boards communities from two different perspectives. First, we investigate frequent strong impacts with respect to each topic in Section 6.3.2. We show that the topic-informed COIN has higher sensitivity and that it provides better interpretability of the cross-community influence relations. Second, in Section 6.3.3 we conduct a follow up analysis that sheds more light onto the impact of the moderating communities on PERSONAL ISSUES that was revealed using the structural COIN in Section 4.2.1 on page 52.

6.3.1 *Determining the Main Community Topic*

As we mentioned in Section 6.2.1, our new notion of topic-informed impact requires us to proceed carefully in selecting only meaningful combinations of topics that we expect to be relevant for each community. In order to find out what topics represent the dominant activity of each forum, we investigate the probability distribution of topics for each community $p(Z|C)$, where Z and C are random variables representing the topics (Z) and communities (C). Even though this distribution is not a direct result of LDA, it can be obtained from the distribution $p(Z|D)$ that is defined by LDA. We define $p(Z|C)$ as the empirical probability that an arbitrary post x contributed to community u was about topic e:

$$p(Z = e|C = u) = \sum_{x \in P_{\cdot u}} p(Z = e|D = x)p(D = x|C = u) \tag{19}$$

$$= \sum_{x \in P_{\cdot u}} p(Z = e|D = x)\frac{1}{|P_{\cdot u}|}, \tag{20}$$

where $P_{\cdot u}$ is the set of all the posts that were contributed to community u.

Assuming that there exists the main topic for each community, our goal is to find a threshold θ representing the minimum probability that is required for a topic to be considered as the main topic of the community. Naturally, if we set θ too low, in addition to the main topics we will cover also many of the secondary or non-representative topics of the community. Conversely, if the threshold θ is too high, the coverage of the main topics will be lower. In order to find the suitable value of θ, we investigated the distribution of the maximum topic probability over the communities. We therefore assume that

the topic that has the maximum probability in a given community represents the dominating activity of the community. We observed that the maximum topic probability was at least 0.2 in 88% of the communities. This means that if we set $\theta = 0.2$, we will cover the main topic for 88% of the communities. This may of course lead to an inclusion of secondary topics if their probability exceeds the threshold. However, this was the case for only 18% of the fora and thus we decided to use $\theta = 0.2$.

6.3.2 *Topic-Informed Pair-Wise Influence Analysis*

The volume of Boards data and the number of communities k (over 600—see Table 1 on page 51) and their topics d (20) makes the analysis of all possible $k(k-1)d^2$ impacts over all the time-windows impractical. Therefore, similarly to our approach in Section 4.2 on page 52 we narrow down the analysis by investigating the *strong impacts* only. Recall that the impact of community u on v is *strong* if it is at least as high as the independence (self-impact) of the community v (see Definition 15 on page 88). By analysing the strong impacts induced by activity associated with only the main community topics we reduced the number of impacts to investigate by more than 1,800-times from 10,530,537 to 5,663.

In this section, we investigate only the impacts $\mathcal{J}_{uevf}$ that were induced by the same topic in both communities, i. e. $e = f$. A strong impact of community u on v thus indicates that the focal members of the community u with respect to the topic highly stimulate activity about the topic in community v. In other words, the community v is dependent on the focal members of u to generate its activity about the topic. Furthermore, in the next Section 6.3.3 we analyse impacts between two communities that were induced by activity related to two different topics.

Topic-Informed Impact is More Sensitive

Due to space reasons, we list only the most frequent strong topic-informed impacts between any two communities with respect to each topic. Table 7 presents the most frequently occurring topic-informed impacts between pairs of communities. The topic labels correspond to Table 6. In order to demonstrate the increased sensitivity of the topic-informed COIN, we contrast the figures that we obtained by applying the structural and the topic-informed models. The column TI of Table 7 contains the number of strong topic-informed impacts between the two communities, whereas the column SI contains the number of strong structural impacts as already presented in Table 3 on page 53. In general, we observe that the values in the TI column

are significantly higher than SI (Wilcoxon signed-rank test, $p = 0.0005$). In the following, we discuss in more detail the top-five most frequent topic-informed impacts from Table 7.

Influence of MODERATORS *on* REPORTED POSTS

We see that similarly to the structural impact, the most frequent impact informed by the topic "moderating" is of MODERATORS on REPORTED POSTS. However, while we observed 41 strong topic-informed impacts in that case, there were 32 strong structural impacts between the two communities. Therefore, by applying the topic-informed COIN we increased the sensitivity to the influence relation that exists with respect to a particular thematic dimension.

Influence of VTFL-ADMIN *on* VTFL-DISCUSSION

Similarly to MODERATORS with respect to the topic "moderating", the community V-TFL ADMIN was observed to strongly impact V-TFL DISCUSSION with respect to "games". While on the one hand the structural COIN was able to reveal this relation, it is very hard to interpret the influence based only on the names of the two fora. On the other hand, the topic-informed impact is not only more sensitive, but it is also easier to interpret. Indeed, the information that the influence between the two communities is related to the activity about "games" was the crucial insight that helped us to understand the relationship between the two communities. We found out that "V-TFL" stands for "Vitality Team Fortress League", a competition for the players of a first person shooting computer game [98]. The role of the admin community was to organise the activity within V-TFL DISCUSSION and in the league in general. We therefore find the high influence of V-TFL ADMIN on V-TFL DISCUSSION intuitive.

Influence of ASK DOCTOR DEMENTO *on* HOLL

The second most frequent strong topic-informed impact reveals dynamics between communities ASK DOCTOR DEMENTO and HOLL that is otherwise left unnoticed using only the structural COIN (16 strong topic-informed impacts vs 2 strong structural impacts). As illustrated in Figure 25 the dominating topic in both of them were "relationships". HOLL (House of Lusty Ladies) is a private community that aims to attract female members who are interested in discussion about relationships and other related topics in often intimate, humorous, and ironic manner. Similarly, ASK DOCTOR DEMENTO (ADD) is a private community whose original purpose was to ridicule some of the topics that are discussed in PERSONAL ISSUES, but over the time ADD be-

topic	community u	community v	TI	SI
moderating	Moderators	Reported Posts	41	29
relationships	Ask Doctor Demento	HoLL	16	2
music	Instruments	Microcube	14	0
games	V-TFL Admin	V-TFL Discussion	13	12
films	Star Trek	Sci-Fi & Fantasy	12	6
religion	Atheism & Agnosticism	Irish Skeptics	11	3
sports	Soccer	Sports	11	4
food	Fitness Logs	Fitness	11	6
fight	Airsoft & Paintball	Airsoft & Paintball Reviews	10	0
school & work	College Work	College Play	10	5
electronics	Tweaking & Modding	Overclocking Logs	9	0
cities	Commuting & Transport	Infrastructure	8	4
money	FS General	FS Sin Bin	8	0
internet	Computers & Technology	Security	7	0

Table 7: The number of strong topic-informed (TI) and structural (SI) impacts of Boards community u on community v with respect to each topic in a descending order of TI. The topic names correspond to the labels as listed in Table 6. We present only strong impacts with at least 7 occurrences, because we observed that less frequent impacts are hard to interpret.

came closer to HoLL. We have already encountered with the community Personal Issues (PI) in Section 4.2.1 on page 52 where we discussed its role as a place for its members to seek advice for their personal problems.

Sometimes the members of ADD refer to HoLL directly expressing their interest in members of HoLL. ADD was mentioned in HoLL 6 times.[2] Conversely, the members of ADD mentioned HoLL 13 times, which indicates an asymmetry in the mutual interest of the two communities. Their close relation is also characterised by another member of ADD who stated that their community is "...like HoLL cept with boyz!...". All of that therefore indicates rich interactions between the two closely related communities that frequently influenced each other, but overall ADD tended to trigger high activity in HoLL more often (16 strong impacts) than the other way around (7 strong impacts).

2 More precisely, it was mentioned *at least* 6 times as we run only a simple fulltext search of "demento". We did not count mentions of abbreviations like "DD" because they are ambiguous.

Influence of INSTRUMENTS *on* MICROCUBE *and Other Influences*

Many of the strong topic-informed impacts in Table 7 were between two communities with a similar name or between two communities whose names suggest a similar focus. This demonstrates that the selection of the main topics we presented in Section 6.3.1 efficiently narrowed down our analysis to only the dominant relations.

For example, the specialised forum about Roland's MICROCUBE, a portable combo speaker intended for use by street performers, seems to be dependent on the activity of users centred around a broader topic—INSTRUMENTS. This suggests that the activity in MICROCUBE relies on the focal members, supposedly musicians, from INSTRUMENTS. However, to confirm this hypothesis a more in-depth qualitative analysis of the text has to be conducted, which is out of the scope of this thesis. Such an analysis is also needed to decide what is the minimum number of strong impacts between two communities to indicate a meaningful influence relation. While it seems that relationships indicated by at least 7 observations of strong topic-informed impacts are meaningful, the less frequent impacts were sometimes difficult to interpret and thus we did not list them in the table.

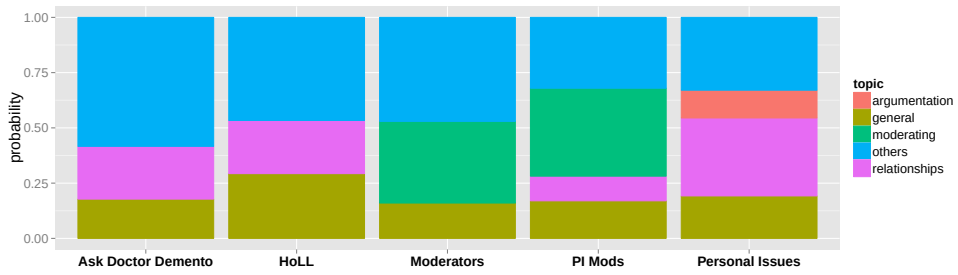


Figure 25: The topic composition of some of the communities that are discussed in the main text. For the sake of brevity, all topics with probability lower than 0.1 are displayed together as "others".

6.3.3 *Topic-Informed Follow-Up Analysis of the Influence of Moderators on Personal Issues*

So far we have required the strong impact to be induced by activity related to the same topic in both communities. In the second part of our analysis, we carefully relax this requirement. As we have already noted in Section 6.2.1, the question of which impact relations induced by a combination of two different topics are meaningful relies on domain knowledge. For example, an influence between two communities indicated by a frequent strong *structural*

impact may be better interpreted by analysing the topics that characterise the activity in the two communities. In order to demonstrate this, we conduct a follow-up analysis of two influence relations that were previously revealed by the structural COIN.

We discussed in Section 4.2.1 on page 52 that two moderating communities from Boards, MODERATORS and PI MODS, influenced PERSONAL ISSUES. We also argued that it is natural, because some discussions in PERSONAL ISSUES attract unhelpful or even mocking behaviour and therefore its activity needs to be regulated. Although these communities are clearly in a relation, Figure 25 shows that the activity in the moderating communities is predominantly associated with the topic "moderating", whereas the members of PERSONAL ISSUES talk mostly about "relationships". This suggests that the moderators discuss the regulation issues in their communities and that the activity they stimulate in PERSONAL ISSUES is mostly associated with the core topics of the community like "relationships". This seems plausible in the light of the previous research that suggests that the regulation in a community is better accepted if it is conducted by authorities whose power is perceived as deserved through e. g. past contribution to the community [53, p. 133–134]. We may thus expect that the moderators frequently contribute to the core topics of PERSONAL ISSUES. However, this is all only a speculation based on the results of the analysis of the structural impact. For that reason, we analysed the topic-informed impact between the fora.

We investigated all strong topic-informed impacts with respect to any combination of the main topics that could explain the relationships between the moderating communities and PERSONAL ISSUES. We found out that the focal members of MODERATORS with respect to "moderating" highly stimulated activity about "relationships" in PERSONAL ISSUES (8 strong impacts). This means that the members of MODERATORS stimulated regularly a high volume of activity in PERSONAL ISSUES about "relationships", while they talked about "moderating" itself in their focal community. We may therefore say that they took an active part in the discussions taking place within PERSONAL ISSUES and only occasionally they facilitated it.

This is further supported by the topic compositions of the two communities illustrated in Figure 25. Whereas the main focus of MODERATORS was naturally "moderating", the main topic of PERSONAL ISSUES were "relationships". However, this was not the case for PI MODS, because in addition to "moderating", its activity was associated with "relationships" as well. We believe that this is the reason why we did not observe any strong topic-informed impact of PI MODS on PERSONAL ISSUES. Since the members of PI MODS discussed "relationships" in both fora, there was no clear focal com-

munity for them with respect to any topic. As a result, we did not gauge a similar influence as the one of MODERATORS.

6.4 CONCLUSION AND DISCUSSION OF THE RESULTS

We have developed an extended and generalised topic-informed COIN that incorporates information about the topics that underpin the activity of the communities. Our motivation was to increase:

- the sensitivity of the cross-community impact and thus to reveal the influence between communities that may be induced by an activity associated with a particular topic;
- the interpretability of the cross-community impact.

In our qualitative study of Boards communities, we have demonstrated that incorporating topics into COIN helps both to reveal and to explain the influence between the communities.

The strong topic-informed impact between two communities was in many cases more frequent than the structural impact. Therefore, the topic-informed COIN offers a higher sensitivity than the structural COIN. However, the fact that many of the influence relations were revealed *already* using structural COIN suggests that both versions of the framework, i. e. topic-informed and structural, yield consistent results. This also means that the purely structural approach is particularly useful whenever the information about the topics is unavailable.

The topic-informed approach improves the interpretability of the cross-community impact. We demonstrated this, for instance, on the follow up analysis of the influence between MODERATORS and PERSONAL ISSUES, and between the communities V-TFL ADMIN and V-TFL DISCUSSION. The insight that the impact of V-TFL ADMIN on V-TFL DISCUSSION is informed by the topic "gaming" was a crucial step in our understanding and interpretation of their relationship. We found out that V-TFL stands for "Vitality Team Fortress League", a computer game competition, and that the role of the admin community was to organise the league and the activity within the V-TFL DISCUSSION community.

While the stratification of the data into topical dimensions often amplified the signal that we were able to extract, other times it inevitably introduced noise into our measurements. A natural remedy is to threshold the signal by e. g. analysing fora only with respect to their main topics. Higher precision of the analysis could perhaps be achieved by using one of the topic-relational models [29, 28] that can jointly infer topics from both the texts of

the documents and the links between then. However, our attempts to calibrate those models to the Boards data has introduced scalability issues. We are still working on overcoming those limits.

The extension of COIN towards multiple dimensions of impact brings novel analytical opportunities. The topical dimensions of impact discussed in this chapter presents one out of several possibilities. Another interesting dimension along which the impact may be stratified is actors' sentiment. Since the sentiment or polarity of the interactions of the actors can affect a community's dynamics [25], e. g. the length of discussion, the addition of the sentiment dimension to COIN may generate novel and interesting insights into cross-community influence. For example, it may shed some light onto whether there are communities that persistently stimulate negative sentiment, and if so, what effects does it have on the influenced communities. The tensor-based notation we adopted promises to enable straightforward integration of such dimensions. Further, the tensor-based notation also promises to allow representation of more complex social interactions in other types of communities than those formed around discussion fora. In the next chapter, we demonstrate this by applying COIN on communities in science, while we leave the other possible extensions of COIN for future work.

7 CROSS-COMMUNITY DYNAMICS IN SCIENCE

Thus far we have focused on the influence between communities of users of online discussion fora leveraging the distributions of their activity and ties that represent their mutual responses. However, as we discussed in Section 2.3 on page 21 the idea of studying the influence between social actors by investigating the responses between their information artifacts was conceived much earlier before the rise of online communities—in the fields of bibliometrics [97] and scientometrics [59]. In particular, large body of research has focused on study of how scientists cite (i. e. respond to) each other in their articles, books, and other information artifacts. Citation analysis have been frequently used for investigation of citation impact and information flow between the individual scientists, i. e. actors (Section 2.3.1 on page 21).

We demonstrate the flexibility of COIN on citation data from communities of computer science researchers. Since majority of the publications in computer science are published at conferences [101], we defined the communities using conferences as a proxy. Hence we use the terms "conference" and "community" interchangeably.

By using COIN, we show how the relationships and *information exchange* [40] a scientific community maintains with other communities contribute to its evolution through its life-cycle in terms of growth, stability, decline, and impact. For example, we expect a successful community to be acknowledged by other communities, i. e. there should be a reasonably high *out-flow* of the information from the community to the other communities. In addition, we expect a successful research community to maintain a sufficient *in-flow* of new knowledge and members. Finally, we expect a specialised community to develop a reasonable level of *introspection* (internal discourse), i.e. its members should be familiar with and refer to the past research outcomes of the community.

We propose that a strong community should keep balance of all the three factors. What is the optimal balance depends on the nature of the community and the stage in its life-cycle. For example, we may expect a new community to be initially less connected to the other communities, while it may get increasingly more cited from outside over the time. Naturally, in other cases a new community may emerge out of already existing communities and therefore may be highly cited immediately after its inception. However, if

The term focal community *is explained in Proposition 1 on page 40.*

the community remains citing mainly itself even in the long term, it may indicate its increasing isolation or even decline. Conversely, a prestigious conference with a broad focus may attract people from disparate disciplines who seek to disseminate their work beyond the boundaries of their focal communities. That is, a community may act as a *hub*. If this is the case, we expect the community to have very high in-flow and out-flow, while the level of its introspection may be lower.

By using COIN, we develop these intuitions into the following *research questions*:

- Can the COIN measures reveal important stages of a community's life-cycle? For example, can they indicate that a community is in decline?
- Can COIN reveal how influential or what role a community has? E. g. can it reveal hub communities that bring together focal members of other communities?

In order to address these questions, we first adapt COIN in order to reflect the characteristics of scientific communities in Section 7.1. After that, we present the data that we analysed in Section 7.2. We proceed with the analysis in two steps. First, we conduct an exploratory analysis that establishes a basis of comparison for the dynamics between the communities in Section 7.3. Second, in Section 7.4 we choose one particular community that appeared to be increasingly isolated from the other communities and we conduct an in-depth qualitative and quantitative analysis seeking to shed more light onto the observed trends. We contrast the results of our analysis with several measures frequently occurring in the literature. Finally, in Section 7.5 we discuss the results and contrast the cross-community approach with a purely introspective approach, i. e. an approach that looks at one community exclusively.

7.1 APPLYING COIN TO RESEARCH COMMUNITIES

We adapt the core definitions of COIN from Chapter 3 in order to cater for the characteristics of scientific communities. Similarly to the discussion communities, we measure the memberships and centrality of each of their members. However, whereas in the discussion communities the links are always between posts belonging to the same forum, a citation from one paper can point to a paper from another conference. Therefore, we ease some of the limits of the data representation that we introduced in Section 3.1 on page 37. Figure 26 illustrates the new data representation that permits a document to refer to another document from a different community. As we explain

in more detail in Section 7.1.1, the new data representation requires us to distinguish the citing and cited community by basing our definitions on the tensor-based notation introduced in Chapter 6. In Section 7.1.2 we discuss the qualitative differences of cross-community impact between discussion communities and the impact between scientific communities. Furthermore, in Section 7.1.3 we adapt the nomenclature of COIN in order to better capture the characteristics of the scientific communities.

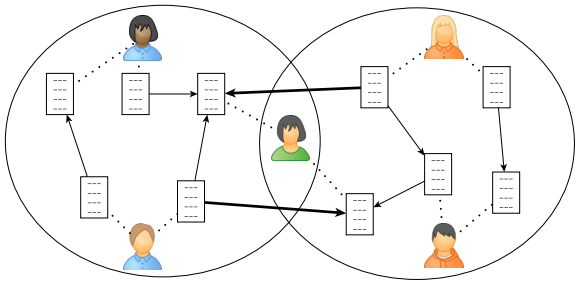


Figure 26: An illustration of a representation of interactions, where a document may be in response to a document from *another* community. A directed link depicts that a document at the source of the link is in *response* to the document at the sink of the link. Responses that are across the communities are emphasised by a bold link. An authorship of a document is denoted by an undirected dotted link.

7.1.1 *Adapting Cross-Community Impact for Scientific Communities*

MEMBERSHIP MATRIX We assume that while a researcher may publish at multiple venues, most of the time there is one field the researcher is mostly focused on [80] which we refer to as her *focal discipline*. Naturally, a researcher may change her focal discipline, but this is unlikely to happen very often (e.g. every year), as it incurs high costs (associated with e.g. learning the state-of-the-art of the new discipline). We therefore expect the focal discipline to remain stable within a time-window whose length is discussed later. A set of authors attending a conference may be perceived as a community corresponding to some (sub-)discipline of science [15]. The distribution of an author's publications over the conferences thus expresses the degree of her membership in each of the communities [80]. We therefore define the $n \times k$ **membership matrix** $\mathbf{M}$ representing a membership of actor i in community u as: $\mathbf{M}_{iu} = |P_{iu}| / \sum_x^k |P_{ix}|$, where P_{iu} is a set of papers contributed by author i to venue u within a time-window.

CENTRALITY TENSOR In order to apply COIN to communities of researchers, it is necessary to measure the centrality of each researcher within each community. As we argued earlier in Section 2.3.1 on page 21, a high number of citations received by a paper corresponds to a high impact it had on the work of other scientists. Although this assumption has been challenged especially when the subjects of the analysis are individual papers or researchers [61, 71, 83], it is believed to be reasonably reliable if the aggregated data are used to compare similar entities (e. g. within the same field) [71, p. 225] and at the highly-cited end of the distribution of citations [83]. A high citation count of a researcher corresponds to a high in-degree (Equation 1 on page 18) in a network of researchers connected by their citations. As with our analysis of Boards.ie, we measure the actor's impact within a community as in-degree. Since a paper may cite another paper from a different community, we measure the actor's centrality with respect to both citing and cited community.

The $n \times k \times k$ **centrality tensor** $\mathcal{C} : \mathcal{C}_{iuv}$ representing a centrality of actor i in community v due to her publications in community u is defined as the total number of citations from papers published at v to the papers published by actor i at conference u. Therefore, the centrality may again be interpreted as a tendency of actor i to stimulate responses from the members of community v. In the case a paper has multiple authors, we assign its citations to each of the co-authors (i. e. we adopt integer counting [71, p. 273]), because in the data available to us there is no quantitative accounting of credit of the individual co-authors [81].

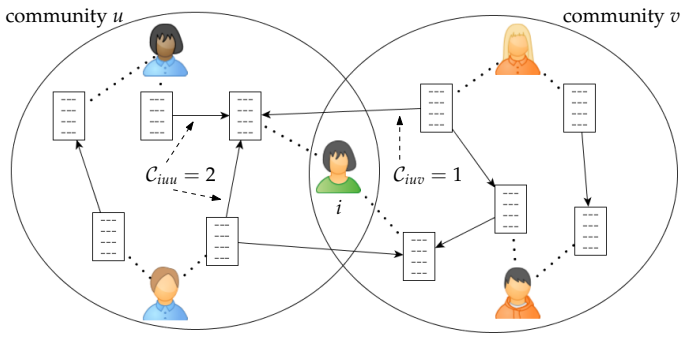


Figure 27: An illustration of the citations-based centrality of actor i (green) at the intersection of the two communities. The undirected dotted links represent authorship, the directed solid links represent citations, and the dashed directed links depict the citations that contribute to the actor's centrality.

Figure 27 illustrates our definition of citations-based centrality. We see that actor i contributed one paper to each of the communities u and v. Because

her paper from community u was cited two times by the other papers from the same community, the centrality of actor i in community u due to her publication in the same community is 2, i. e. $\mathcal{C}_{iuu} = 2$. Analogously, since the same publication by actor i from community u received one citation from a paper from community v, the centrality of actor i in the other community v due to her publication in u is 1, i. e. $\mathcal{C}_{iuv} = 1$.

IMPACT MATRIX Analogously to Definition 6 on page 43, the cross-community impact $\mathbf{J}_{uv}$ of community u on v can be obtained as:

$$\mathbf{J}_{uv} = \frac{\sum_x^n \mathbf{M}_{xu}\mathcal{C}_{xuv}}{\mathbf{s}_u} \tag{21}$$

As before $\mathbf{s}$ is a vector of community sizes. Alternatively, using the contracted tensor product introduced in Definition 13 on page 83, the cross-community **impact matrix** can be obtained as: $\mathbf{J} = \mathbf{S}^{-1}(\mathcal{M} \otimes_2 \mathcal{C})$, where $\mathbf{S}$ is a diagonal matrix of sizes as in Definition 6 on page 43 and $\mathcal{M}$ is a $n \times k \times k$ tensor:

$$\mathcal{M}_{iuv} = \begin{cases} \mathbf{M}_{iu} & \text{if } u = v \\ 0 & \text{otherwise} \end{cases}$$

We may interpret the impact from u to v as a tendency of the members of v to cite the members of u. Furthermore, as we discussed in Section 2.3 on page 21 citations may be interpreted as an indicator of (reverse) information flow as illustrated by Figure 28. Therefore the impact of u to v may also be interpreted as a measure of *information flow* from u to v.

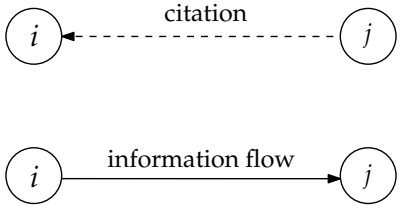


Figure 28: Illustration of the relation between information flow and citations. A citation from actor j to actor i indicates an explicit engagement of actor j with a document from actor i. It suggests that actor j ingested the content of the document and thus that some information "flowed" in the opposite direction from actor i to j.

7.1.2 *Cross-Community Impact In Discussion vs Scientific Communities*

There are a few important differences between the cross-community impact that we analysed in discussion fora and the impact between scientific communities. One difference is that the impact between two fora was induced by the activity of their overlapping members (see Section 3.1 on page 37), whereas in scientific communities a member of one community may cite a member of another community directly, i. e. they do not have to share the two communities. In addition, while a post in a discussion community can be in reply to exactly one post, a paper usually cites multiple other papers. Furthermore, posts in discussion communities are usually shorter than research papers. The amount of information within a paper is thus generally higher than the amount within a post. Therefore, there may be also a greater flow of information between papers than between posts. Indeed, as we discussed in Section 2.3.1 on page 21, Dietz, Bickel, and Sheffer [28] were able to estimate the information flow between research papers using a topic-relational model. While a reply between two users of discussion fora sometimes coincides with flow of information (see Section 4.2.1 on page 52), we interpreted it primarily as an indicator of the activity stimulation. In contrast with that, since citations are more abundant, may occur directly between members of different communities, and relate documents whose length is longer we believe that they are more likely to indicate flow of information.

7.1.3 *Aggregate Measures as Applied to Scientific Communities*

Apart from the notion of cross-community impact, we also introduced its aggregate measures of importance (Equation 6) and dependence (Equation 7). Importance measured the total impact a community had on other communities and dependence quantified the total impact other communities had on the community. We also defined the importance and dependence entropy as a measure of their heterogeneity (Equations 8 and 9). While these terms are suitable for discussion communities, they are liable to be misunderstood when applied in the context of scientific communities. In order to avoid possible misconceptions, we adopt different terminology.

As we noted earlier, the impact may be interpreted as an information flow between two communities and thus the aggregated measures indicate the overall flow from and to the community:

OUT-FLOW The *out-flow* of a community is the total impact the community has on other communities (Equation 6 on page 44).

OUT-FLOW ENTROPY The *out-flow entropy* of a community quantifies the heterogeneity of the out-flow, i. e. to how many distinct communities the information flows from the community (Equation 8 on page 45).

IN-FLOW The *in-flow* of a community is a sum of all the impacts other communities have on the community (Equation 7 on page 44).

IN-FLOW ENTROPY The *in-flow entropy* of a community measures from how many distinct communities the information flows to the community (Equation 9 on page 45).

INTROSPECTION Since the impact $\mathbf{J}_{uu}$ measures the tendency of members of community u to self-cite the same community we call the impact *introspection*.

In short, we adapted the core definitions of COIN to reflect the specifics of citation networks and communities in science. Before we present the results of our analysis, we discuss the choice and preparation of the data we used for the analysis.

7.2 ARNETCITE DATA-SET

Our background in computing suggested that we could interpret more easily results from research of communities in computer science. We therefore computed the COIN measures for a broad range of venues in computer science. After which we narrowed our focus to a subset of the venues that are related to Artificial Intelligence (AI).

We focused on AI as we are familiar with the main paradigms and events within it. We adopted the definitions of the sub-fields of computer science by Martins et al. [65], who proposed a ranked list of computer science conferences, Perfil-CC. The ranking was obtained by a poll of domain experts and was also shown to be in correspondence with another popular ranked list CORE (Computing Research Association of Australasia) [65, 26]. Each conference is classified into one of the groups A, B, and C according to their presumed merit with A representing the top-tier venues. Since it is a recommended practise in bibliometrics to compare only authors or their groups that are reasonably similar to each other (Section 2.3 on page 21), the Perfil-CC classification also gives us a basis for such comparison. Even though Perfil-CC distinguishes Machine Learning (ML) from AI, we decided to merge the two sub-fields as ML is indeed a sub-discipline of AI. This results into 87 different AI conferences in total. For the space reasons, we list the conferences in Table 12 on page 128.

7.2.1 *Enriching Arnet Data With Citations from CiteSeerX*

In Section 2.3.3 on page 26 we presented two bibliographic databases for computer science: DBLP [58] and CiteSeerX [60] (previously known as CiteSeer [38]). Although DBLP contains manually curated high quality metadata for a broad range of venues in computer science, it contains only little citation data. Opposite to that, CiteSeerX is an autonomous citation indexer that contains more citation data but also more noise. Apart from that, both data-bases suffer the author name ambiguity problem, i.e. that two or more authors can have the same name. Tang et. al [95] proposed a machine learning model to deal with the name ambiguity and integrated the DBLP data with other available metadata. The resulting database, ArnetMiner, combines and cleans data from ACM Digital Library, CiteSeer, DBLP, and the Web.

For all the reasons discussed above, we chose ArnetMiner data-set from September 2013 for our analysis [94]. Despite the fact that ArnetMiner contains generally high quality metadata, we found out that not all citations between the documents indexed by CiteSeerX are contained in ArnetMiner data-set. Therefore, we further copied the missing citations from CiteSeerX data from August 2011.[1] We refer to the resulting data-set as *ArnetCite*.

In order to assure good quality of ArnetCite, we further carried out several data-cleaning operations. These were namely to guarantee that all citations are among the indexed documents; that there is not obviously wrong meta-data (e.g. erroneous year of publishing like "0"); or to assign a common name to venues that have been renamed or merged [36]. For example, as discussed later in Section 7.4.5, the European Workshop on Case-Based Reasoning was renamed to the European Conference on Case-Based Reasoning and later merged with the International Conference on Case-Based Reasoning. Hence, in order to properly trace the group of scientists in case-based reasoning research, we merged those venues into one Case-Based Reasoning (CBR) community. See Appendix A for the full description of the preparation of ArnetCite. The cleaned and integrated data is also available online [12].

# papers	# authors	# citations	# venues
1,246,455	766,293	2,281,946	5,224

Table 8: Elementary statistics of the analysed part of ArnetCite.

While ArnetCite indexes a substantial number of papers up to year 2012, it covers a reasonable number of citations only up to year 2009. This is due to

1 Our attempts to download a more up-to-date version of the data via OAI were unsuccessful due to service instability.

an earlier version of CiteSeerX data as one of the sources of the citation data. Since the number of citations drops sharply after year 2009, we cannot be sure that the data for that year is representative. Likewise, there is low number of records in ArnetCite prior 1990. Therefore, we used only the citation data between 1990 and 2008 in our analysis. Out of the 87 AI conferences listed by Perfil-CC, we found 59 in ArnetCite that are listed in Table 12 for space reasons. This corresponds to a coverage of nearly 68%. Table 8 lists some elementary statistics of the data-set.

7.2.2 *Data Segmentation*

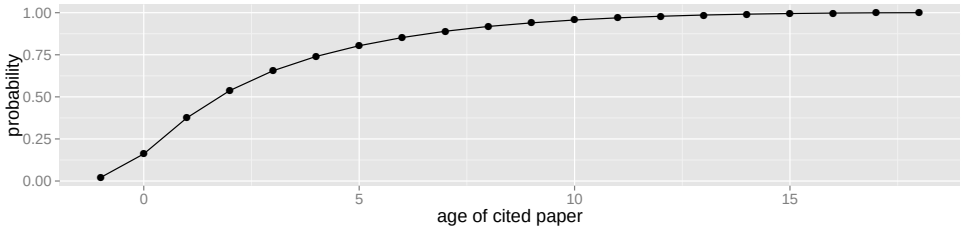


Figure 29: The cumulative probability distribution function of the age of a cited paper. Please note that the age of a cited paper can be negative. That happens when a paper cites another paper that is expected to be published in some of the subsequent years. See Appendix A for more details.

As we did in the analysis of discussion communities in Chapters 4 and 5, we segment the data using a sliding time-window. However, whereas it takes usually only weeks or months for the majority of the replies between discussion communities to occur (see Section 4.1 on page 51), scientific discourse operates at a much slower pace. For a paper to be cited, typically another author has to read it, reference it from a new paper that is then peer-reviewed and published. Further, some of the conferences are held biennially or even triennially, and therefore the window should be wide enough to capture them [65]. Since none of the AI conferences we focused on in our analysis was held triennially, but some were biennial (e.g. IJCAI), we chose a window of 4 years overlapping by 3 years. This guarantees that the window covers at least two occurrences of each conference. As Figure 29 depicts, the median age of a cited paper in ArnetCite is 2 years and the window of 4 years covers nearly 75% of all the citations.

7.3 CROSS-COMMUNITY ANALYSIS OF AI CONFERENCES

We applied COIN in an analysis of the AI communities in order to reveal important changes in their life-cycles and to investigate the relationships a community maintains with other communities.

Earlier in the chapter, we explained how out-flow, in-flow, and introspection characterise interactions between communities. In particular, we argued that high in-flow or introspection may uncover different types of communities such as hub or isolated communities. However, since research communities operate generally differently depending on their culture [11], what is an extreme value for one community may be a "norm" for other community. For example, an applied research community may be less cited from other communities, i. e. may have a low out-flow, compared with a pure research community, because the applied research may be more valuable outside science, e. g. in industry. In order to establish a basis for comparison for our analysis, we therefore analyse only the AI communities, which we assume to be reasonably similar with respect to their publication and citation practises.

We use the aggregate measures of information flow to investigate the trends that may indicate important changes in the community's life-cycle in terms of its impact on other communities. We expect high *out-flow* and *introspection* to be a sign of a strong community, because the outcomes of its strong internal discourse (introspection) are acknowledged by the other communities (out-flow). We may say that such communities are "exporters" [40]. Additionally, a high *in-flow* and low *introspection* of a community indicates that the community acts as a *hub* that brings together researchers from diverse communities. Since those researchers are likely to cite the papers published in their focal communities, we expect a high in-flow to the hub community. Finally, we expect a community that is growing increasingly *introspective* and isolated (low *out-flow* and *in-flow*) to experience a decline, because it indicates that the community is unable to attract the interest of new researchers or other communities—it becomes self-referential.

In order to investigate the main trends of the aggregate measures, we divided the data into 3 periods: *early* period between the years 1990–1996; *middle* period between 1997–2002; and *late* period covering the years 2003–2008. Figure 30 depicts the mean *out-flow* (x-axis) and *introspection* (y-axis) of the AI communities in each period. For the sake of brevity, we discuss only the six communities listed in Table 9. We chose those communities because their high values of one or more of the aggregate measures suggest that they represent characteristic examples of the types of communities that we described above, i. e. hubs, "exporters", or self-referential communities.

However, we encourage an interested reader to consult the online version of the plot that allows an interactive analysis of changes on an annual basis [12]. In Table 10 we list the mean values of flow statistics for each community along with the community size measured as the cardinality of the set of its members (see Section 3.3 on page 42).

community	COLT	NIPS	IJCAI	ICML	ILP	CBR
class	A	A	A	A	A	B

Table 9: The AI conferences that we analysed along with their classes.

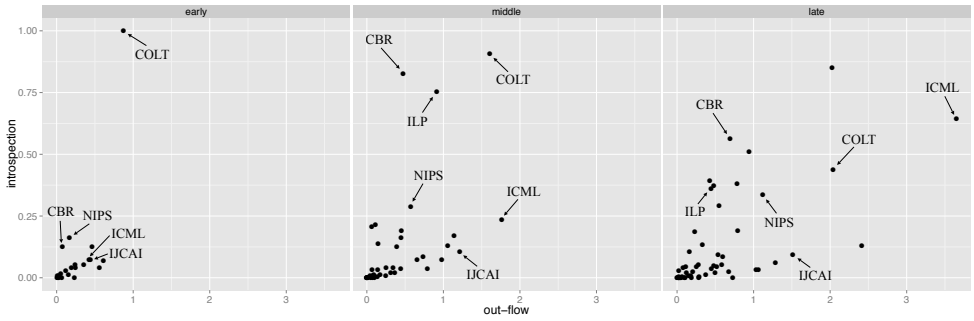


Figure 30: The mean out-flow and introspection for each AI community in early, middle, and late periods. For the sake of brevity, only the communities that are discussed in the main text are annotated. Please note that ICML was very close to IJCAI in the early period because we measured very similar introspection and out-flow for both of them. ILP is not depicted in the early period, because it appears in ArnetCite for the first time in 1997.

COLT In the early period, we see that the community COLT (Annual Conference on Computational Learning Theory) had both very high introspection and out-flow. This suggests that the community was strong as it maintained high level of both internal discourse but at the same time its outcomes were referenced from the outside. Over the time, the out-flow of COLT has increased substantially while its introspection lowered. Together with the fact that Perfil-CC ranked COLT as a class A conference, it indicates that COLT has evolved from a relatively highly self-referential community into a more open community, while it has increased its already high impact on the other communities.

NIPS Another community with a relatively high introspection in the early period was NIPS (CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS). While maintaining high level of internal discourse over the time, the out-flow from NIPS has increased considerably in the middle and especially in the late period. As its name suggests, NIPS started as a conference with predominantly computational neuroscience focus. Over the time, however, it became one of the major venues in machine learning, artificial intelligence, and statistics. This transition to a more open conference with a broader focus is also indicated by a rise of the community's in-flow (see Table 10). We may therefore conjuncture, that the increase of its total impact can be attributed to the successful transition from a small but strong community with a narrow focus to an open, yet still strong community maintaining a sufficient level of internal discourse. Similarly to COLT, NIPS is listed by Perfil-CC as a class A conference and our observations support that indeed it has evolved into a strong and highly influential community.

IJCAI One of the most respected conferences in AI is the INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE (IJCAI) that has been held biennially since 1969 [46]. As a rather large conference with a broad focus covering many sub-disciplines of artificial intelligence, IJCAI is a typical *hub* venue, where many researchers and practitioners from various fields and of various background meet. It is therefore no surprise that it is characterised by relatively small introspection, but very high in-flow (see Table 10). This indicates that IJCAI attracts researchers, who publish frequently in other communities with a perhaps narrower focus that corresponds to their domain of expertise, and who seek to disseminate the results of their work beyond the boundaries of their focal communities.

ICML ICML (INTERNATIONAL CONFERENCE ON MACHINE LEARNING) is a premium machine learning conference (class A in Perfil-CC). In contrast with IJCAI, its higher introspection suggests that it is less of a hub and that its attendees tend to regularly publish their work at it. The more than 8 times increase of its out-flow between the early and late periods suggests a rising interest in and consumption of the machine learning methods within other communities. ICML thus became an "exporter" of the machine learning techniques. This may correspond to a recent information explosion that lead some researchers and practitioners to talk about the era of "Big Data" [14].

CBR Although the CASE-BASED REASONING (CBR) community started, similarly to NIPS, with a high introspection, in contrast with NIPS the CBR community remained highly introspective (i. e. self-referential) also in the

community	out-flow	introspection	int./out-flow	in-flow	size	period
CBR	0.08	0.13	**1.64**	0.39	**88**	early
CBR	0.47	0.83	**1.74**	0.91	**152**	middle
CBR	0.69	0.56	**0.81**	1.07	**119**	late
COLT	0.87	1.00	1.15	1.86	71	early
COLT	1.61	0.91	0.57	1.15	48	middle
COLT	2.04	0.44	0.21	0.60	55	late
ICML	0.42	0.08	0.18	0.65	66	early
ICML	1.76	0.24	0.13	0.96	139	middle
ICML	3.65	0.64	0.18	3.12	210	late
IJCAI	0.44	0.07	0.16	**5.60**	305	early
IJCAI	1.22	0.11	0.09	**4.60**	231	middle
IJCAI	1.51	0.09	0.06	**7.01**	282	late
ILP	0.92	0.75	**0.82**	0.23	38	middle
ILP	0.44	0.36	**0.82**	0.68	51	late
NIPS	0.17	0.16	0.97	**0.03**	537	early
NIPS	0.57	0.29	0.50	**1.08**	447	middle
NIPS	1.12	0.34	0.30	**2.55**	498	late

Table 10: Mean aggregate measures of the AI conferences that are discussed in the main text. The "int./out-flow" presents the ratio of introspection and out-flow. The figures mentioned in the text are in bold. The size is rounded to integers for the sake of brevity.

middle and late periods. We observed a similar trend of a high introspection relative to the out-flow also for the ILP community (INTERNATIONAL CONFERENCE ON INDUCTIVE LOGIC PROGRAMMING). The high introspection relative to the out-flow (see Table 10) suggests that the community was unable to attract new researchers and broader interest in its topics. Another reason may be that the community simply reached the limits of its paradigm [54].

In either case, such dynamics could naturally lead to a gradual decline of the community in terms of size and impact. For instance, its members may find it difficult to access resources like funding for their research, because grant applications are frequently assessed by their peers who may come from different communities. Furthermore, few researchers would choose to enter a community that seems to be increasingly isolated from the discourse of other communities, because the isolated community is less visible to them (and thus they may not even know about it), or because they may

be concerned about its prospects in the future. Indeed, we observed that the size of the CBR community first grew from 88 in the early period to 152, but then it lowered to 119.

An analysis and explanation of these trends is of crucial importance to many stakeholders of scientific communities, because it may help them to understand at what stage of its life-cycle their community is and thus to make better-informed decisions. Therefore, in order to shed some light on these trends, we take the CBR as a subject of a more in-depth analysis in the next section, while we leave the ILP community for future work.

7.4 RISE AND FALL OF THE CASE-BASED REASONING PARADIGM

In the previous section we analysed six AI communities and chose the CBR community for a closer investigation because the combination of the low out-flow, high introspection, and lowering size suggests its decline. While there may be many reasons why a community's size is lowering, we believe that an increasing isolation of a community may be an important factor that contributes to the community's decline. Therefore, we investigate the relations of CBR with the other communities using the aggregate measures of in-flow, out-flow, and their entropy (see Section 7.1.3) over the time. Our aim is to answer the following *questions*:

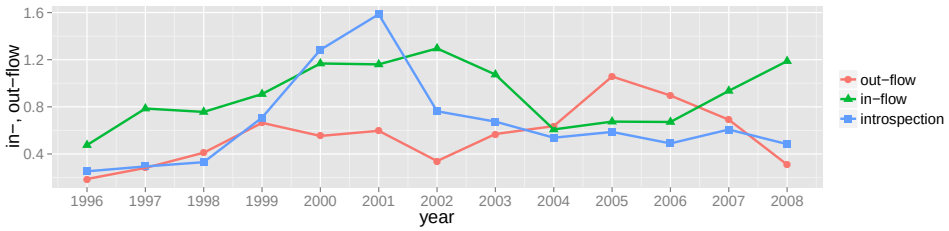
- How did the relations of CBR with the other communities evolve over the time? Did it become increasingly isolated?
- With how many communities did CBR maintain relationships? Did CBR have a narrow focus?
- Was the community in decline in terms of its size, number of publications, and citation impact? If so, since when?

We validate our findings by qualitative analysis of the history of the CBR community and by contrasting the observed trends of the COIN measures with other measures frequently occurring in the literature. In order to refer to time consistently for all the measures further in the text, we refer to a time-window $[t, t']$ only by its end year t'. For example, the values for the year 1996 were measured in the window [1993, 1996].

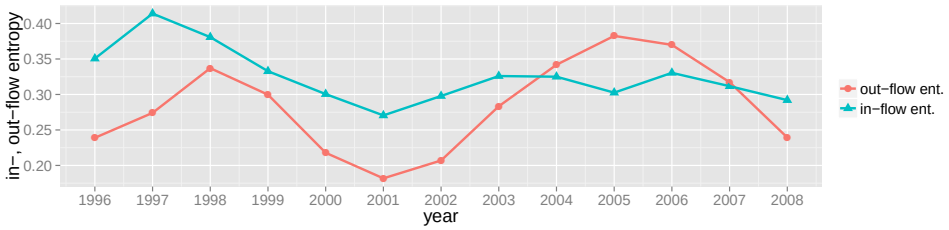
7.4.1 CBR *Was Increasingly Isolated*

As we saw in the previous section, the CBR community featured very high level of introspection, especially relatively to its moderate out-flow. Figure 31a depicts the change of the aggregate measures over the time. We see that since

the early beginnings of the community, its introspection was steadily rising up to the year 2001, when it peaked and subsequently lowered to a stable level between the years 2002–2008. Similarly to that, the in-flow reached its peak in 2002, then it was lowering until the year 2006, when, however, the trend reversed again. In contrast with the introspection and in-flow, the community had relatively low out-flow, which means that it was referencing a lot itself and other communities, but those communities did not reference it back to such an extent.



(a) In-flow, out-flow, and introspection of CBR per time-window.

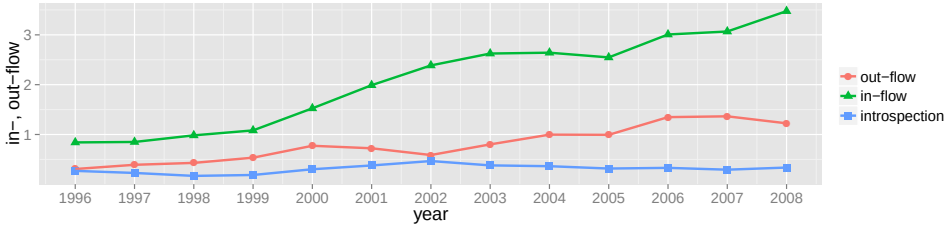


(b) Entropy of in-flow and out-flow of CBR.

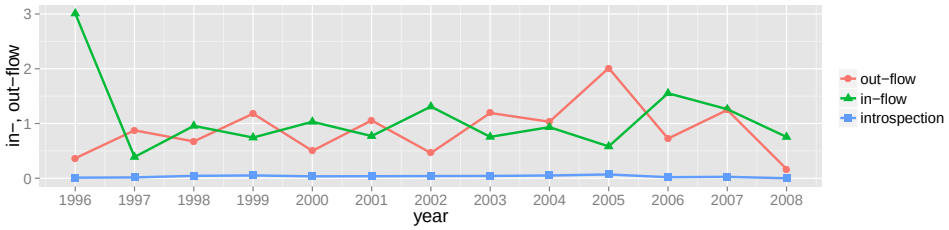
Figure 31: In-flow, out-flow, and their heterogeneity (entropy) for the CBR community.

This contrasts with the trends that we commonly observed for the other communities. As two examples, we include the trends of a class A conference, NIPS, in Figure 32a and a class B conference, JELIA (European Conference on Logics in Artificial Intelligence), in Figure 32b [48]. The other plots are available online [12]. We see that in both cases the introspection is relatively low compared to in-flow and out-flow. Furthermore, in both cases the trends of in-flow and out-flow are not as concerning as in the case of CBR. The in-flow and out-flow of JELIA were relatively in balance. Even though the in-flow of NIPS was higher than its out-flow, the both flows were growing and thus it rather indicates the openness of the community as we discussed in Section 7.3. In contrast, the high in-flow and introspection of

CBR relative to its lower out-low indicates that this community was indeed increasingly isolated compared to the other communities.



(a) In-flow, out-flow, and introspection of NIPS per time-window.



(b) In-flow, out-flow, and introspection of JELIA per time-window.

Figure 32: Aggregate measures of in-flow and out-flow for the NIPS and JELIA communities.

The out-flow of CBR reached its peak in 2005 and then it began to fall rather sharply. We can also see that the out-flow was generally lagging behind the introspection by approximately 4 years. The lag can be explained by the necessity to first develop ready-to-use solutions before they can be used by (or "exported" to) other communities. The high out-flow in 2005 was therefore likely induced by the research outputs from the beginning of the time-window [2002, 2005]. This suggests, together with the peak in introspection in 2001, that the community reached its climax around the years 2001–2002. The strong discourse of the community in that time resulted in a high impact on the other communities. However, the very high level of introspection accompanied by comparatively much lower out-flow between 1999–2002 could detract new potential researchers from joining that community. Indeed, few researches would choose to enter a field that seems to be increasingly isolated from the rest of the scientific discourse.

7.4.2 CBR *Had a Narrow Focus*

This is further supported by Figure 31b illustrating the change of entropy, i. e. heterogeneity, of the in-flow and out-flow of CBR. In the very beginning, the CBR was influenced by many other communities. For a young paradigm, this might be expected as it is still yet to develop its own discourse. Since the peak in 1997, its in-flow entropy was steadily decreasing until 2001. Similarly to that, the heterogeneity of its out-flow was increasing at first, but then the out-flow entropy reached its bottom in 2001. The dips in 2001 can be explained by the very high introspection in that period. It means that the majority of the citation activity was fuelled by the internal discourse. In the Kuhnian terms [54], we may say that the paradigm reached the climax of its articulation in around 2001. After that, both in-flow and out-flow entropy grew until 2005–2006, since when they lowered again. This indicates a gradually narrower focus of the community in the last years of our data.

7.4.3 *The Member Base of* CBR *Was Rigid*

We believe that one of the factors that contributed to the high introspection and narrow focus of CBR was a low influx of new members to the community. In order to show that CBR was indeed more rigid in terms of its member base, we run another experiment. Our hypothesis was that if the CBR community was attracting new researchers less frequently than the other communities, its member base should be more stable in time than expected. We quantified the stability of the member base of a community as its self-similarity measured by Jaccard index. The key challenge was to define what is "expected". In our experiment, we decided to compare the stability of CBR member base with the rest of the communities that were also classified as class B conferences by Perfil-CC.

For each time-window between 1993–2008, we measured for each AI community u a Jaccard similarity of the fuzzy sets of its members in time-window t and subsequent time-window $t+1$:

$$\mathrm{js}(\mathbf{M}^t_{\cdot u}, \mathbf{M}^{t+1}_{\cdot u}) = \frac{\sum_x^n \min(\mathbf{M}^t_{xu}, \mathbf{M}^{t+1}_{xu})}{\sum_x^n \max(\mathbf{M}^t_{xu}, \mathbf{M}^{t+1}_{xu})}, \tag{22}$$

where $\mathbf{M}^t_{\cdot u}$ represents the fuzzy set of the members of community u at time-window t (Section 3.3 on page 42). Equation 22 is a generalisation of a common Jaccard index defining similarity of two crisp sets as a ratio of the cardinalities of their intersection and union. A standard way to obtain an intersection of two fuzzy sets is to include each element with the minimum membership in the two sets [110]. Analogously, a union is obtained

as a maximum of the two membership values [110]. As already discussed in Section 3.3, a cardinality of a fuzzy set can be defined as a sum of the memberships of its elements.

We computed an expected value of the Jaccard similarity for each window excluding the values of CBR. This way we obtained a paired sample of two time-series: one for the "average" class B conference and one for the CBR community itself. The values are listed in Table 11. It turned out that CBR had significantly higher self-similarity than the rest of the communities (Wilcoxon signed-rank test, $p = 0.02$).

year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
CBR	0.67	0.56	0.65	0.58	0.66	0.52	0.54	0.55	0.58	0.57	0.58	0.57	0.51
class B	0.57	0.46	0.34	0.55	0.46	0.62	0.43	0.61	0.46	0.47	0.47	0.58	0.54

Table 11: Jaccard similarity of CBR and of the rest of the class B conferences over the time.

7.4.4 *Main Findings of the Cross-Community Analysis of* CBR

Based on our cross-community influence analysis we may say that the community was increasingly isolated, had a narrow focus, and its member base was rigid and shrinking. Therefore, we believe that after the period of an initial growth, the community was in decline since approximately year 2001. In order to validate our observations, we investigate the history of CBR along with a few alternative performance measures.

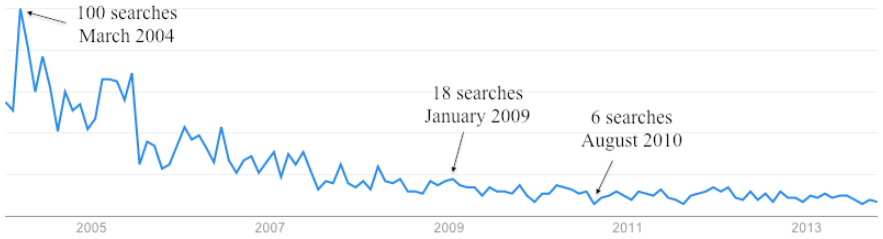
7.4.5 *History of* CBR

Case-based reasoning is an artificial intelligence paradigm that emerged out of Cognitive Science research. It "solves new problems by retrieving stored records of prior problem-solving episodes (cases) and adapting their solutions to fit new circumstances" [56]. At the time of inception, it was a novel approach to many problems within AI that promised to provide a new perspective on the development of intelligent systems. Encouraged by those promises, special tracks devoted to case-based reasoning were organised at the top-tier artificial conference IJCAI in the years '97, '99, '01, '03, and '05 (IJCAI is a biennial conference). However, since 2007 there was no such track at IJCAI [47].

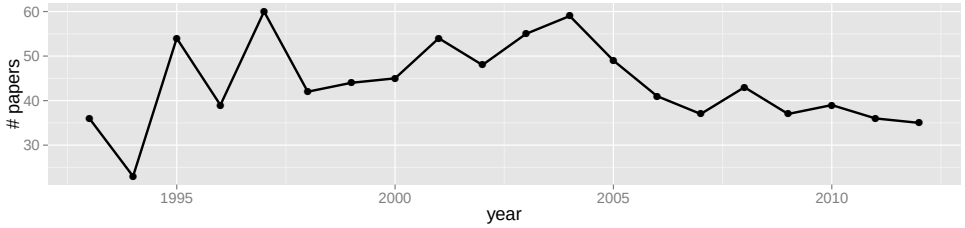
At the same time, a core research community has formed around specific workshops and conferences. European Workshop on Cased-Based Reasoning (EWCBR) was first held in 1993 and then it was transformed into

a biennial EUROPEAN CONFERENCE ON CASED-BASED REASONING (ECCBR) in 2002. INTERNATIONAL CONFERENCE ON CASED-BASED REASONING (ICCBR) was organised biennially since 1995 and then, after a merger with ECCBR, annually since 2009.

In short, the case-based reasoning paradigm evolved from a rise accompanied by an establishment of specific conferences and recognition at a general and prestigious AI forum into a merger of its main venues and a less prominent presence at IJCAI.



(a) number of Google searches



(b) number of papers

Figure 33: The trends of interest in CBR.

7.4.6 *Decay of the Output of and Interest in* CBR

This suggests that the interest in case-based reasoning research was growing until approximately the year 2005—the time of the last observed special track at IJCAI. Figure 33 depicts two trends that shed some light on the changing interest in case-based reasoning over time. The top Figure 33a illustrates the number of searches of the phrase "case based reasoning"[2] the users of Google Search performed between January 2004 and September 2013 as obtained from the Google Trends [21].[3] We see that the trend peaks in March 2004 when 100 searches were performed. Since then, the number

2 The trend using the "case-based reasoning" phrase looks similar.

3 Data prior 2004 are unavailable due to the limitations of the service.

has been diminishing to about 10 searches per month from 2009 onwards. We note that the falling number of Google searches coincides with the fall of out-flow illustrated by Figure 31a. The bottom Figure 33b shows the total number of regular papers published at CBR per year. The highest number of papers was in the years 1997 (60 papers) and 2004 (59 papers). Then it lowered to 35–43 between 2006–2012. These figures suggest that the interest in case-based reasoning in general and the research output of the CBR community in particular have been declining since the year 2005. However, quantity of the research output does not necessarily correlate with its impact [83]. For that reason, we investigated four other statistics.

7.4.7 *Citation Impact and Other Performance Measures of* CBR

Figure 34 illustrates the trends of the four additional statistics per each time-window: PageRank (PR, Equation 4 on page 28), group in-degree (GI, Equation 3 on page 18), 3-years conference impact factor (CIF) [65], and size measured as the cardinality of the set of the community members (see Section 3.3 on page 42). PageRank was computed on a graph of communities, i.e. a network in which two communities are connected if there exist one or more citations between the papers published within the communities. Group in-degree is a total number of citations received by the papers published by CBR from papers published elsewhere. Conference impact factor (CIF) of a conference in year t is the average number of citations a paper published by the community within $[t-3, t-1]$ received from all the papers published at t. Please recall that we refer to a time-window only by its end year.

All measures except CIF were decreasing since 2005. CIF does not indicate any clear trend, but is in direct contrast with the other measures—especially with GI. While GI was falling since 2005, we observed a rather moderate rise of CIF since that year. Since CIF includes self-citations whereas GI does not, the rise of CIF can be attributed to the rising introspection of CBR (see Section 7.4.1).

7.4.8 *Conclusion*

In conclusion, we observed on many different scales an initial rise of CBR followed by its subsequent decline. The rise was characterised by a growing ability of the community to attract new members, to adopt research outputs from other communities, and to develop its own internal discourse. The community probably reached its zenith around the years 2001-2002, when the outcomes of the strong discourse of the community induced a high impact on the other communities in the later years.

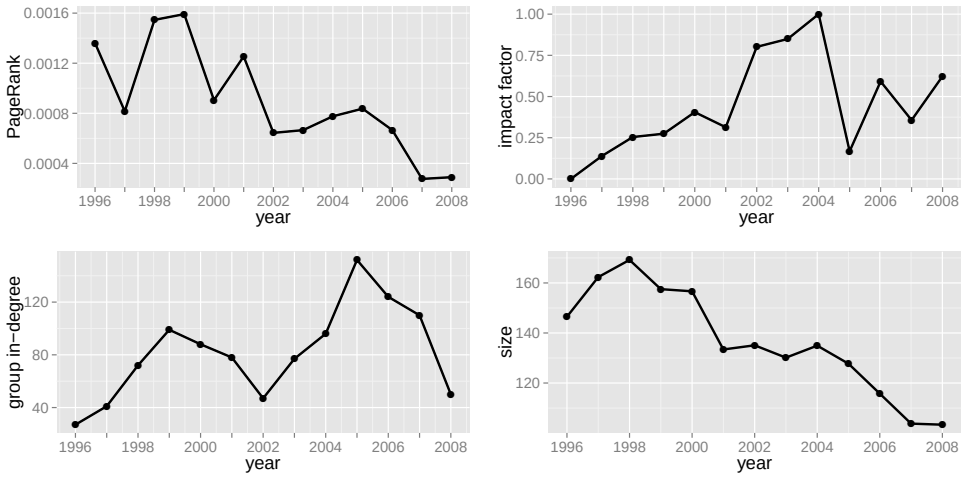


Figure 34: PageRank, 3-years conference impact factor, group in-degree, and size of CBR over time.

Even though the number of publications was moderately rising even after 2001, we tend to believe that this was an effect of "inertia". It indeed requires some time for the researchers to recognise that the community has already passed its zenith. Therefore, after the short period of "inertia", the community started to shrink and became more influenced by the other communities, whereas those communities did not referenced it back to the same extent.

Moreover, while there were special interest tracks dedicated to the case-based reasoning paradigm at IJCAI, a top-tier venue in AI, there has not been any in the recent years. Since 2004, the output measured as the number of papers of the community was decreasing, while we observed a high degree of self-citation. Even though our purely structural approach to the analysis cannot provide exhaustive explanation to the causes driving this dynamics, based on our results we believe that one of them was an inability of the community to attract new members.

7.5 DISCUSSION OF THE RESULTS

We have demonstrated the flexibility of COIN on cross-community analysis of scientific communities. Although we analysed only a subset of AI communities, we expect COIN to be generally applicable to other research communities. For example, we believe that many interesting insights may be obtained by applying COIN on data from semantic web, business processes, description logic, and other related communities.

The results of our analysis showed that COIN is suitable for revealing and explanation of the relations between the communities. For instance, we were able to identify hub communities that bring together researchers from different fields. Furthermore, we showed that the COIN measures generate valuable insights into how communities evolve through their life-cycles. In particular, we observed that a very high introspection and in-flow in combination with low out-flow and narrow focus of a community indicates a state of isolation that may lead to the community's decline in terms of size and citation impact. We believe that this was the case of the CBR community that evolved from the period of an initial growth to its zenith in approximately 2001, after which the community declined.

Introspective Analysis May Not Be Enough

These conclusions contrasts with an introspective analysis conducted by the CBR community itself [41]. The authors conducted a bibliometric analysis of ECCBR up to the year 2008.[4] They observed a regular rise of new topics within the community and suggested that it "can be considered a sign of a healthy research area" [41]. However, the introspective approach to their analysis bears one risk: while a community may indeed develop regularly new topics and abandon the old ones, it is a necessary but not sufficient condition for a community to be "healthy". A community may, for instance, become gradually isolated, i. e. not being cited from the other communities, yet still it may change the themes of its discourse. In such situation, however, the community members may find it increasingly challenging to obtain external support for their research, e. g. research grants, because their work may not be recognised by their peers from other disciplines.

Handling Self-citations

How to handle an author's self-citations has been a persistent topic in bibliometrics since the introduction of citation indexes in science [35, 71, 86, 5]. Our analysis of the CBR community showed that self-citations can be very valuable for the investigation of the community's dynamics, but also that they should be handled with care. The results can radically differ for measures of the global impact like the conference impact factor (CIF) on the one hand and the cross-community measures of COIN on the other hand. In spite of the rising CIF of CBR towards the end of our data, we argued that in fact the community experienced decline indicated by measures on multi-

4 Recall that ECCBR stands for European Conference on Case-Based Reasoning. See Section 7.4.5 for more details.

ple different scales. As in other applications of bibliometrics, investigation of multiple indicators is therefore appropriate [81, 63, 18].

Peer Review May Not Be Enough

Furthermore, it is often recommended to combine the indicators with peer-review [81, 71]. However, while peer-review proved its efficacy for identifying promising new areas and researchers, it may be unsuitable for dividing the established areas into those that are flourishing and those that are not [63]. For example, the senior scientists involved in peer-review may be influenced by their past impressions of who the best performers are, but those impressions may become obsolete over time [81]. Quantitative analysis of cross-community relations as enabled by COIN may thus help to overcome these biases and to recognise the actual state of the communities.

COIN as a Bibliometric Tool

In order to interpret the results, a familiarity with the field is as important as other factors influencing the conclusions, such as accuracy and completeness of the data. Different indicators may be susceptible to different biases and thus it is only their combination, especially when using incomplete data like ArnetCite, that promises to yield more accurate insights. In this regard, we believe that COIN is a valuable piece in the mosaic of already existing bibliometric methods and that it provides valuable insights into the evolution of social dynamics of science from the cross-community perspective.

8 CONCLUSIONS

We have demonstrated that communities fundamentally shape the effects of influence between the individual actors: e. g. information is shared with the whole discussion community and *not* with the individual members. Furthermore, communities are often in a relationship whereby one community strongly affects other communities, e. g. moderating or administrating communities in discussion fora by definition control the dynamics of the other communities. The ability to quantify, analyse, and interpret cross-community influence is therefore essential for research and exploitation of influence between individual actors wherever their communities affect how the actors interact.

In this thesis, we have developed a structural approach to cross-community influence that fills the gap in quantification, analysis, and explanation of influence relations between dynamic social communities. We believe that the exploitation of the influence on the community level will become common in the future, e. g. for efficient information dissemination where the information is shared with the whole community, or for monitoring and predictive analytics of communities.

8.1 SUMMARY OF THE THESIS

We have developed a computational model for cross-community influence, COIN. We demonstrated the flexibility and efficacy of our structural approach on a range of qualitative studies and simulation experiments. Furthermore, we have extended the purely structural model to handle topics.

INFLUENCE MEASUREMENT AND ANALYSIS COIN enabled us to reveal the existence of the influence between three different types of communities: general-purpose discussion fora; question-answering communities; and communities of computer science researchers. Our analysis showed a wealth of diverse community influence relations and interactions, such as a rise of global authorities; the changing patterns of influence experienced by a particular community; emergence of communities with broad topics playing the role of hubs; or increasing isolation and drop in influence of scientific communities. We believe that the insights generated by COIN may help the communities' stakeholders to better understand or manage their communi-

ties. For example, a community that grows increasingly isolated from other communities is at risk of becoming irrelevant and may disappear due to less external opportunities. As COIN offers a range of indicators that can detect such dynamics, it may be used by the stakeholders as a basis for decision making or predictive analytics. Indeed, these reasons motivated the integration of COIN into the PULSAR analytical platform from SAP [70].

INFORMATION DIFFUSION Another application of cross-community influence and COIN is to enable efficient information diffusion. In many situations the community is the receiver of information, and not just its individual members. We have extended previously defined models of information diffusion to the community level. Although we proved that the maximisation problem under the extended models is NP-hard, there are efficient ways how to tackle it by leveraging heuristic approaches. Namely a COIN-derived heuristic, *impact focus*, led to high user and community adoptions for both static and dynamic social networks.

TOPICAL DIMENSIONS OF INFLUENCE Although we primarily focused on purely structural analysis, we demonstrated the extensibility of COIN by generalising its core measures to capture topics that may underpin the observed influence between communities. While the structural approach is useful if the information about topics is unavailable due to e. g. legal or technical reasons, the integration of the topics increased the signal we were able to extract from the data and improved the interpretability of our analysis.

8.2 LIMITATIONS OF THE THESIS AND DIRECTIONS FOR FUTURE RESEARCH

INFLUENCE MEASUREMENT Even though we generally observed a strong relation between cross-community impact and influence between communities, the impact may not measure the influence accurately. For example, we observed that focal members of influential communities disseminate information within influenced communities. However, social influence and actors' homophily are generally confounded in social networks [89], and thus the high adoption of information by influenced communities could be partially induced by external factors. An important theme for future research is thus an experimental measurement of cross-community influence in controlled studies [10]. Further, as the level of cross-posting activity in SAP was generally lower than in Boards, we also observed less influence between the SAP communities. However, that does not mean necessarily that there indeed was less influence, because the suitability of the measures we pro-

posed is commensurate to the amount of signal in the data. Therefore, if the signal is lower, as in the case of SAP, the quantification and tracking of cross-community influence remains a challenge. We believe that a higher sensitivity may be achieved by a graphical model that would generate the observed network of actors and documents from a *latent* probability distribution representing the cross-community influence relations.

INFORMATION DIFFUSION We believe that the impact focus strategy for selection of seed communities can be further improved by penalising overlap between selected seed communities. The improved strategy would be biased towards targeting communities that do not overlap and therefore are more likely to represent different parts of the network. This principle was successfully used for the actor seed selection problem [23]. Furthermore, even though we observed the cross-community impact to be correlated with the measures of language diffusion, more research is needed to validate the extended models. For example, Saito et al. [87] proposed an expectation maximisation approach for parameter estimation of actor-level diffusion models based on empirical cascades data. One direction for future research is thus an extension of their approach to cross-community diffusion models.

SENTIMENT DIMENSIONS OF INFLUENCE Another possible extension of this work is to correlate polarity or sentiment of community's content with cross-community influence. Chmiel et al. [25] demonstrated how sentiment in online communities affects a community's dynamics, e.g. the length of discussions, and proposed that the sentiment analysis may help the stakeholders to keep their community alive. Since an influential community may have both negative or positive influence, sentiment analysis of cross-community influence could therefore help the stakeholders to understand sensitivities associated with particular topics or behaviours, e. g. spamming.

THEORY OF CROSS-COMMUNITY INFLUENCE While some trends, like the rise of hub communities, were similar in all analysed systems, others were unique to a particular type of communities. More research is thus necessary to determine which cross-community influence phenomena are typical for which class of communities. Since we were able to apply COIN to three different types of communities, we believe that COIN may be applicable to other community types such as company teams communicating via email. Systematic analysis of cross-community influence on other data-sets or on other types of communities than that were analysed in this thesis may eventually lead to a theory of cross-community influence. In this regard, we consider COIN as a significant step towards such an endeavour.

A

ARNETCITE DATA PREPARATION

This appendix details the preparation of the ArnetCite data introduced in Section 7.2 on page 105. We used the data in order to analyse the AI communities listed in Table 12. Section A.1 describes how we cleaned the Arnet data. After that, we describe the integration of the Arnet data with the CiteSeerX citation data into the ArnetCite data-set in Section A.2.

A.1 CLEANING OF ARNET DATA

In order to assure the quality of our data, we carried out the following data-cleaning operations:

1. We deleted 14,043 papers with no known venue.
2. We deleted all authors with an empty name and we deleted 17,343 papers of those authors.
3. We deleted 60,596 citations that related at least one paper that was not contained in Arnet, i. e. citations pointing outside of the data-set.
4. We deleted 3,774 authors that did not author any paper or whose paper was deleted in one of the earlier steps.
5. We deleted 42 venues without any paper.
6. We deleted 42,232 citations that pointed to a paper that was published longer than 1 year *after* the citing paper, because we considered the citations pointing very long into the future as erroneous.
7. Since Arnet also contains some books, we deleted all "venues" that occurred in one year only.

A.2 DATA INTEGRATION OF ARNET WITH CITESEERX

After we cleaned the Arnet data, we copied additional citations between the papers in Arnet from CiteSeerX as we noted earlier in Section 7.2 on page 105. In order to copy the citations, we attempted to match each of the 2,243,965 papers in Arnet to each of the 9,219,151 papers indexed by

CiteSeerX.[1] We did the matching between an Arnet paper x and a CiteSeerX paper y by using a heuristic based on their titles and the lengths of the titles l_x and l_y. Before the matching, we removed any non-alphanumerical characters from the titles and converted the titles into a plain sequence of lower-case words separated by a single space. We deemed the two papers identical:

1. either if the titles of the two papers were identical;
2. or if all of the following held true:
   a) the lengths differed, say $l_x < l_y$;
   b) the shorter title was fully contained in the longer title;
   c) $\frac{1-(l_y-l_x)}{l_y} \geq \theta$, where θ is a threshold.

We introduced the second condition to account for the cases when the two titles were practically the same but one of them contained a small error. For example, we matched these two non-identical titles: "system test cost modelling based on event rate analysis" and "*a* system test cost modelling based on event rate analysis" (note the additional "a" in the beginning of the second title). We tried different values of the threshold and found out that the suitable value is 0.7. This value implies that the lengths of the two titles do not differ more than by 30%. In total, we were able to match 628,295 papers. As a result, we copied additional 1,248,346 citations from CiteSeerX.

Table 12: The AI conferences listed by Perfil-CC along with the status of their coverage by ArnetCite.

name	full name	class	status
AAAI	AAAI Conference on Artificial Intelligence	A	✗
AAMAS	International Joint Conference on Autonomous Agents and Multiagents Systems	A	✓
ABS	Agent-based Simulation Workshop	C	✗
AIA	Artificial Intelligence and Applications Conference	B	✗
AIAI	IFIP International Conference on Artificial Intelligence Applications and Innovations	A	✓
AIED	International Conference on Artificial Intelligence in Education	B	✓

Continued on the next page

1 Please note that these figures correspond to the *total* number of papers in the data-sets. As we described in Section 7.2, we analysed only a subset of the papers.

Table 12 – *Continued from the previous page*

name	full name	class	status
AIL	International Conference on Artificial Intelligence and Law	C	×
AIME	Conference on Artificial Intelligence in Medicine in Europe	C	✓
AIPS	Conference on Artificial Intelligence Planning Systems	B	✓
AISAT	International Conference on Artificial Intelligence in Science and Technology	C	×
AISC	International Conference on Artificial Intelligence and Symbolic Computing	B	✓
ALAMAS	European Symposium on Adaptive Learning Agents and Multi-Agent Systems	B	×
ALT	International Conferences on Algorithmic Learning Theory	B	✓
AMAI	International Symposium on Artificial Intelligence and Mathematics	B	✓
ANIREM	Workshop on Agents, Norms and Institutions for Regulated Multiagent Systems	C	×
ANTS	International Workshop on Ant Colony Optimization and Swarm Intelligence	B	✓
AOIS	Agent-Oriented Information Systems Workshop	C	✓
CAIA	Conference on Artificial Intelligence for Applications	B	×
CEC	IEEE Congress on Evolutionary Computation	A	✓
CIA	International Workshop on Cooperative Information Agents	B	✓
CIMSA	IEEE International Conference on Computational Intelligence for Measurement Systems and Applications	B	×
CIRAS	International Conference on Computational Intelligence, Robotics and Autonomous Systems	B	×
COIN	Workshop on Coordination, Organization, Institutions and Norms in Agent Systems	B	✓
COLT	Annual Conference on Computational Learning Theory	A	✓
CogSci	Annual Conference of the Cognitive Science Society	A	×
DALT	International Workshop on Declarative Agent Languages and Technologies	C	✓

Continued on the next page

Table 12 – *Continued from the previous page*

name	full name	class	status
DS	International Conference on Discovery Science	C	✓
DSS	IFIP International Conference on Decision Support Systems	B	×
E4MAS	International Workshop on Environments for Multiagent Systems	B	✓
ECAI	European Conference on Artificial Intelligence	A	✓
ECAIM	European Conference on Artificial Intelligence in Medicine	B	×
ECML	European Conference on Machine Learning	A	✓
EKAW	International Conference on Knowledge Engineering and Knowledge Management	A	✓
EKM	European Conference on Knowledge Management	B	×
ESANN	European Symposium on Artificial Neural Networks	B	✓
EUMAS	European Workshop on Multi-Agent Systems	B	✓
EuroCOLT	European Conference on Computational Learning Theory	B	✓
EuroGP	European Conference on Genetic Programming	B	✓
FAABS	IEEE Workshop on Formal Approaches to Agent-Based Systems	B	✓
FOCI	IEEE Symposium on Foundations of Computational Intelligence	A	✓
FSKD	International Conference on Fuzzy Systems and Knowledge Discovery	B	✓
FUZZ	IEEE International Conference on Fuzzy Systems	A	✓
GECCO	Genetic and Evolutionary Computation Conference	A	✓
HIS	International Conference on Hybrid Intelligent Systems	A	✓
IAAI	Conference on Innovative Applications in Artificial Intelligence	A	×
IAT	ACM International Conference on Intelligent Agent Technology	A	✓
IBERAMIA	Ibero-American Artificial Intelligence Conference	A	✓
ICAI	International Conference on Artificial Intelligence	B	×
ICANN	International Conference on Artificial Neural Networks	A	✓

Continued on the next page

Table 12 – *Continued from the previous page*

name	full name	class	status
ICAPS	International Conference on Automated Planning and Scheduling	B	✓
ICCB	International Conference on Case-Based Reasoning	B	✓
ICCI	IEEE International Conference on Cognitive Informatics	B	✓
ICGA	International Conference on Genetic Algorithms	B	✓
ICIL	International Conference on Intelligent Systems	C	×
ICML	International Conference on Machine Learning	A	✓
ICMLA	International Conference on Machine Learning and Applications	B	✓
ICMLC	International Conference on Machine Learning and Cybernetics	C	✓
ICNC	International Conference on Natural Computation	B	✓
ICONIP	International Conference on Neural Information Processing	A	×
ICPR	International Conference on Pattern Recognition	A	✓
ICTAI	IEEE International Conference on Tools with Artificial Intelligence	A	✓
IDA	International Symposium on Intelligent Data Analysis	A	✓
IDEAL	International Conference on Intelligent Data Engineering and Automated Learning	C	✓
IEEEIS	IEEE Conference On Intelligent Systems	A	×
IFIP AI	IFIP Artificial Intelligence	A	✓
IFSA	International Fuzzy Systems Association World Congress	A	×
IJCAI	International Joint Conference on Artificial Intelligence	A	✓
IJCNN	IEEE International Joint Conference on Neural Networks	A	×
ILP	International Conference on Inductive Logic Programming	A	✓
IPMU	International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems	B	✓
IPS	IEEE International Conference on Intelligent Processing Systems	B	×

Continued on the next page

Table 12 – *Continued from the previous page*

name	full name	class	status
ISDA	International Conference on Intelligent Systems Design and Applications	B	✓
ITS	International Conference on Intelligent Tutoring Systems	B	×
IWANN	International Work-Conference on Artificial and Natural Neural Networks	C	✓
JELIA	European Conference on Logics in Artificial Intelligence	B	✓
KES	International Conference on Knowledge-Based and Intelligent Information and Engineering Systems	B	✓
MABS	International Workshop on Multi-Agent-Based Simulation	B	✓
MCS	International Workshop on Multiple Classifier Systems	B	✓
MLDM	IAPR International Conference on Machine Learning and Data Mining	B	✓
MLMTA	International Conference on Machine Learning and Applications	C	✓
MLSP	IEEE International Workshop on Machine Learning for Signal Processing	B	×
NAFIPS	North American Fuzzy Information Processing Society International Conference	B	×
NIPS	Neural Information Processing Systems	A	✓
SEAL	International Conference on Simulated Evolution and Learning	B	×
UAI	Conference in Uncertainty in Artifical Intelligence	A	✓
WAIS	International Workshop on Artificial Intelligence and Statistics	C	×
WCCI	IEEE World Congress on Computational Intelligence	A	×

BIBLIOGRAPHY

[1] After Hours. Boards.ie Forum After Hours. http://www.boards.ie/vbulletin/forumdisplay.php?f=7. Accessed: 2013-12-13. (Cited on page 53.)

[2] Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010. (Cited on pages 15 and 34.)

[3] Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639, 2011. (Cited on pages 3, 30, and 32.)

[4] Jaime Arguello, Brian Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling Ling, and Xiaoquing Wang. Talk to me: Foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2006. (Cited on page 30.)

[5] Douglas N. Arnold and Kristine K. Fowler. Nefarious numbers. *Notices of the AMS*, 58(3):434–437, 2011. (Cited on page 120.)

[6] Sitaram Asur, Srinivasan Parthasarathy, and Doygu Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(4):16, 2009. (Cited on page 33.)

[7] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'06)*. ACM, 2006. (Cited on page 30.)

[8] Lars Backstrom, Ravi Kumar, Cameron Marlow, Jasmine Novak, and Andrew Tomkins. Preferential behavior in online groups. In *Proceedings of the international conference on Web search and web data mining (WSDM'08)*, pages 117–128. ACM, 2008. (Cited on pages 30 and 31.)

[9] Brett W. Bader and Tamara G. Kolda. Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software (TOMS)*, 32(4):635–653, 2006. (Cited on pages 82 and 83.)

[10] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the international conference on World Wide Web (WWW'12)*. ACM, 2012. (Cited on pages 3, 30, 33, and 124.)

[11] Tony Becher and Paul Trowler. *Academic tribes and territories: Intellectual enquiry and the culture of disciplines*. McGraw-Hill International, 2001. (Cited on page 108.)

[12] Václav Belák. A Structural Approach to Community-level Social Influence Analysis—Online Supplementary Material. http://belak.net/doc/2014/thesis.html, 2014. (Cited on pages 7, 77, 78, 90, 106, 109, and 113.)

[13] Václav Belák, Marcel Karnstedt, and Conor Hayes. Life-cycles and mutual effects of scientific communities. *Procedia—Social and Behavioral Sciences*, 22:37–48, 2011. (Cited on pages 25 and 26.)

[14] Big Data. What is big data? http://strata.oreilly.com/2012/01/what-is-big-data.html. Accessed: 2013-11-01. (Cited on page 110.)

[15] Maria Biryukov and Cailing Dong. Analysis of computer science communities based on DBLP. In *Research and advanced technology for digital libraries*, pages 228–235. Springer, 2010. (Cited on pages 25, 26, and 101.)

[16] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003. (Cited on page 88.)

[17] Boards. Boards.ie—Now Ye're Talkin'. http://www.boards.ie. Accessed: 2013-12-13. (Cited on pages 5 and 49.)

[18] Johan Bollen, Herbert Van de Sompel, Aric Hagberg, and Ryan Chute. A principal component analysis of 39 scientific impact measures. *PloS one*, 4(6):e6022, 2009. (Cited on pages 22 and 121.)

[19] Francesco Bonchi. Influence propagation in social networks: A data mining perspective. *The IEEE Intelligent Informatics Bulletin*, 12(1), 2011. (Cited on pages 32 and 33.)

[20] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 30(1): 107–117, 1998. (Cited on page 28.)

[21] CBR-Trends. Google Trends of "case based reasoning". https://www.google.com/trends/explore?q=case+based+reasoning. Accessed: 2013-10-09. (Cited on page 117.)

[22] Jonathan Chang and David M Blei. Relational topic models for document networks. In *International Conference on Artificial Intelligence and Statistics*, 2009. (Cited on page 88.)

[23] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'09)*. ACM, 2009. (Cited on pages 32, 33, 68, 69, 73, 74, 80, and 125.)

[24] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'10)*, pages 1029–1038. ACM, 2010. (Cited on pages 26, 28, 29, 33, and 68.)

[25] Anna Chmiel, Julian Sienkiewicz, Mike Thelwall, Georgios Paltoglou, Kevan Buckley, Arvid Kappas, and Janusz A Hołyst. Collective emotions online and their influence on community life. *PloS one*, 6(7), 2011. (Cited on pages 98 and 125.)

[26] CORE. Computing Research and Education Association of Australasia. http://www.core.edu.au/. Accessed: 2013-11-15. (Cited on page 105.)

[27] Derek J. de Solla Price. Networks of Scientific Papers. *Science*, 146 (3683):510–515, 1965. (Cited on page 21.)

[28] Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the International Conference on Machine learning (ICML'07)*. ACM, 2007. (Cited on pages 23, 88, 97, and 104.)

[29] Laura Dietz, Ben Gamari, John Guiver, Edward Snelson, and Ralf Herbrich. De-Layering Social Networks by Shared Tastes of Friendships. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'12)*, 2012. (Cited on pages 88 and 97.)

[30] Milad Eftekhar, Yashar Ganjali, and Nick Koudas. Information cascade at group scale. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'13)*. ACM, 2013. (Cited on pages 26, 34, 35, and 37.)

[31] Martin G. Everett and Stephen P. Borgatti. The centrality of groups and classes. *Journal of Mathematical Sociology*, 23(3):181–201, 1999. (Cited on pages 18 and 35.)

[32] Dalibor Fiala. Mining citation information from CiteSeer data. *Scientometrics*, 86(3):553–562, 2011. (Cited on page 27.)

[33] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486 (3-5):75–174, 2010. (Cited on pages 13, 14, and 42.)

[34] Noah E. Friedkin. *A structural theory of social influence*. Cambridge University Press, 1998. ISBN 0521454824. (Cited on pages 3, 17, 19, 35, and 37.)

[35] Eugene Garfield. Citation indexes for science. *Science*, 144:649–54, 1964. (Cited on pages 21 and 120.)

[36] Eugene Garfield et al. Citation analysis as a tool in journal evaluation. *Science*, 178:471–479, 1972. (Cited on pages 22 and 106.)

[37] Nancy L. Geller. On the citation influence methodology of Pinski and Narin. *Information Processing & Management*, 14(2):93–95, 1978. (Cited on page 23.)

[38] C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. CiteSeer: An automatic citation indexing system. In *Proceedings of the ACM conference on Digital libraries*. ACM, 1998. (Cited on pages 21, 27, and 106.)

[39] Markus Gmür. Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, 57(1):27–57, 2003. (Cited on page 24.)

[40] Robert L. Goldstone and Loet Leydesdorff. The import and export of cognitive science. *Cognitive Science*, 30(6):983–993, 2006. (Cited on pages 24, 25, 99, and 108.)

[41] Derek Greene, Jill Freyne, Barry Smyth, and Pádraig Cunningham. An analysis of research themes in the CBR conference literature. In *Advances in Case-Based Reasoning*, pages 18–43. Springer, 2008. (Cited on page 120.)

[42] Steve Gregory. Fuzzy overlapping communities in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011, 2011. (Cited on pages 14, 15, 34, and 42.)

[43] Adrien Guille, Hakim Hacid, Cécile Favre, and Djamel A. Zighed. Information Diffusion in Online Social Networks: A Survey. *SIGMOD Record*, 42(2):17, 2013. (Cited on pages 17, 32, and 33.)

[44] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999. (Cited on page 88.)

[45] David Huffaker. Dimensions of leadership and social influence in online communities. *Human Communication Research*, 36(4):593–617, 2010. (Cited on pages 30, 32, 50, 55, and 56.)

[46] IJCAI—About. International Joint Conference on Artificial Intelligence—About. http://ijcai.org/aboutijcai.php. Accessed: 2013-10-23. (Cited on page 110.)

[47] IJCAI—Past. International Joint Conference on Artificial Intelligence—Past Conferences. http://ijcai.org/past/index.php. Accessed: 2013-10-23. (Cited on page 116.)

[48] JELIA. JELIA—European Conference on Logics in Artificial Intelligence. http://www.jelia.eu/. Accessed: 2013-12-29. (Cited on page 113.)

[49] Benjamin F. Jones, Stefan Wuchty, and Brian Uzzi. Multi-university research teams: shifting impact, geography, and stratification in science. *Science*, 322(5905):1259–1262, 2008. (Cited on page 24.)

[50] Marcel Karnstedt, Tara Hennessy, Jeffrey Chan, Partha Basuchowdhuri, Conor Hayes, and Thorsten Strufe. Churn in social networks. In *Handbook of Social Network Technologies and Applications*, pages 185–220. Springer, 2010. (Cited on page 66.)

[51] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'03)*, pages 137–146. ACM, 2003. (Cited on pages 3, 17, 26, 32, 33, 50, 68, 69, 71, 72, 73, 74, and 75.)

[52] Jon Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999. (Cited on pages 5 and 29.)

[53] Robert Kraut and Paul Resnick. *Building successful online communities*. The MIT Press, 2012. (Cited on pages 3, 4, 13, 30, 66, 85, and 96.)

[54] Thomas S. Kuhn. *The structure of scientific revolutions*. University of Chicago press, 1996. (Cited on pages 111 and 115.)

[55] Theodoros Lappas, Kun Liu, and Evimaria Terzi. A survey of algorithms and systems for expert location in social networks. In *Social Network Data Analytics*, pages 215–241. Springer, 2011. (Cited on pages 28 and 29.)

[56] David B. Leake. Case-based reasoning. In *Encyclopedia of Computer Science*, pages 196–197. John Wiley and Sons Ltd., Chichester, UK, 2003. ISBN 0-470-86412-5. (Cited on page 116.)

[57] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'07)*, pages 420–429. ACM, 2007. (Cited on pages 32, 33, and 73.)

[58] Michael Ley. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *String Processing and Information Retrieval*, pages 1–10. Springer, 2002. (Cited on pages 26, 34, and 106.)

[59] Loet Leydesdorff and Staša Milojević. *Scientometrics*, chapter International Encyclopedia of Social and Behavioral Sciences. Elsevier, 2015. Available online: http://arxiv.org/abs/1208.4566. (Cited on pages 8, 21, and 99.)

[60] Huajing Li, Isaac Councill, Wang-Chien Lee, and C. Lee Giles. CiteSeerX: an architecture and web service design for an academic document search engine. In *Proceedings of the International Conference on World Wide Web (WWW'06)*, pages 883–884. ACM, 2006. (Cited on pages 27 and 106.)

[61] Michael H MacRoberts and Barbara R MacRoberts. Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, 40(5):342–349, 1989. (Cited on pages 23 and 102.)

[62] Jeffrey Mann, Tom Austin, Nikos Drakos, Carol Rozwell, and Andrew Walls. Predicts 2013: Social and Collaboration Go Deeper and Wider. https://www.gartner.com/doc/2254316/predicts--social-collaboration-deeper, 2012. Accessed: 2013-12-16. (Cited on page 3.)

[63] Ben R. Martin. The use of multiple indicators in the assessment of basic research. *Scientometrics*, 36(3):343–362, 1996. (Cited on pages 21, 22, and 121.)

[64] Ben R. Martin and John Irvine. Assessing basic research: some partial indicators of scientific progress in radio astronomy. *Research policy*, 12 (2):61–90, 1983. (Cited on page 21.)

[65] Waister Silva Martins, Marcos André Gonçalves, Alberto H.F. Laender, and Nivio Ziviani. Assessing the quality of scientific conferences based on bibliographic citations. *Scientometrics*, 83(1):133–155, 2010. (Cited on pages 23, 105, 107, and 118.)

[66] Jiří Matoušek and Jaroslav Nešetřil. *Invitation to discrete mathematics*. Oxford University Press, 2008. (Cited on page 12.)

[67] Andrew K. McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002. (Cited on pages 88 and 90.)

[68] M. McGlohon and M. Hurst. Community structure and information flow in USENET: Improving analysis with a thread ownership model. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'09)*. AAAI, 2009. (Cited on pages 15, 30, 31, 34, 37, 41, 47, and 48.)

[69] Yasir Mehmood, Nicola Barbieri, Francesco Bonchi, and Antti Ukkonen. CSI: Community-Level Social Influence Analysis. In *Machine Learning and Knowledge Discovery in Databases*, pages 48–63. Springer, 2013. (Cited on pages 34, 35, and 37.)

[70] Adrian Mocan. ROBUST Community Management Support for the SCN. http://scn.sap.com/community/research/blog/2012/11/16/robust-community-management-solutions-for-the-scn, 2012. Accessed: 2013-12-16. (Cited on pages 8, 65, and 124.)

[71] Henk F. Moed. *Citation analysis in research evaluation*, volume 9. Springer, 2005. (Cited on pages 4, 21, 22, 23, 24, 26, 48, 102, 120, and 121.)

[72] Sergio Lopez Montolio, David Dominguez-Sal, and Josep Lluis Larriba-Pey. Research Endogamy as an Indicator of Conference Quality. *SIGMOD Record*, 42(2):11, 2013. (Cited on page 24.)

[73] MySQL. MySQL 5.1 Reference Manual—Full-Text Stopwords. http://dev.mysql.com/doc/refman/5.1/en/fulltext-stopwords.html. Accessed: 2013-11-07. (Cited on pages 56 and 89.)

[74] George Nemhauser, Laurence Wolsey, and Marshall Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978. (Cited on page 73.)

[75] Christoph Neuhaus and Hans-Dieter Daniel. Data sources for performing citation analysis: an overview. *Journal of Documentation*, 64(2): 193–210, 2008. (Cited on pages 21 and 26.)

[76] Mark E. J. Newman. *Networks: an introduction*. Oxford University Press, 2010. ISBN 0199206651. (Cited on pages 11, 18, 21, and 28.)

[77] Official Google Blog. We knew the web was big... http://googleblog.blogspot.ie/2008/07/we-knew-web-was-big.html, 2008. Accessed: 2013-11-28. (Cited on page 28.)

[78] OpenNLP. Apache OpenNLP version 1.5.0. http://opennlp.apache.org. Accessed: 2013-11-07. (Cited on pages 56 and 89.)

[79] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005. (Cited on pages 15 and 43.)

[80] Akshay Patil, Juan Liu, and Jie Gao. Predicting group stability in online social networks. In *Proceedings of the international conference on World Wide Web (WWW'13)*. ACM, 2013. (Cited on pages 15, 25, 26, 31, 34, 37, 40, 47, and 101.)

[81] David A. Pendlebury. The use and misuse of journal metrics and other citation indicators. *Archivum immunologiae et therapiae experimentalis*, 57 (1):1–11, 2009. (Cited on pages 22, 23, 102, and 121.)

[82] Vaclav Petricek, Ingemar J Cox, Hui Han, Isaac G. Councill, and C. Lee Giles. A comparison of on-line computer science citation databases. In *Research and Advanced Technology for Digital Libraries*, pages 438–449. Springer, 2005. (Cited on page 27.)

[83] T.J. Phelan. A compendium of issues for citation analysis. *Scientometrics*, 45(1):117–136, 1999. (Cited on pages 102 and 118.)

[84] Gabriel Pinski and Francis Narin. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12(5):297–312, 1976. (Cited on pages 23 and 28.)

[85] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3): 130–137, 1980. (Cited on pages 56 and 89.)

[86] Jan Reedijk and Henk F. Moed. Is the impact of journal impact factors decreasing? *Journal of Documentation*, 64(2):183–192, 2008. (Cited on pages 23 and 120.)

[87] Kazumi Saito, Masahiro Kimura, Kouzou Ohara, and Hiroshi Motoda. Selecting information diffusion models over social networks for behavioral analysis. In *Machine Learning and Knowledge Discovery in Databases*, pages 180–195. Springer, 2010. (Cited on pages 80 and 125.)

[88] SAP. SAP Community Network. `http://scn.sap.com`. Accessed: 2013-12-13. (Cited on pages 5 and 49.)

[89] Cosma Rohilla Shalizi and Andrew C Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011. (Cited on page 124.)

[90] Xiaolin Shi, Jun Zhu, Rui Cai, and Lei Zhang. User grouping behavior in online forums. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'09)*. ACM, 2009. (Cited on page 52.)

[91] Myra Spiliopoulou. *Social Network Data Analytics*, chapter Evolution in Social Networks: A Survey, pages 149–175. Springer, 2011. (Cited on pages 3, 13, and 20.)

[92] Jimeng Sun and Jie Tang. *Social Network Data Analytics*, chapter A survey of models and algorithms for social influence analysis, pages 177–214. Springer, 2011. (Cited on pages 3, 17, 33, 50, and 75.)

[93] Tao Sun, Wei Chen, Zhenming Liu, Yajun Wang, Xiaorui Sun, Ming Zhang, and Chin-Yew Lin. Participation maximization based on social influence in online discussion forums. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'11)*. AAAI, 2011. (Cited on page 30.)

[94] Jie Tang. ArnetMiner Citation Network Dataset. http://arnetminer.org/lab-datasets/citation/DBLP_citation_Sep_2013.rar, 2014. Accessed: 2014-1-15. (Cited on page 106.)

[95] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: extraction and mining of academic social networks. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008. (Cited on page 106.)

[96] The Thunderdome. Boards.ie forum the thunderdome. http://www.boards.ie/vbulletin/forumdisplay.php?f=484. Accessed: 2013-12-13. (Cited on page 53.)

[97] Mike Thelwall. Bibliometrics to webometrics. *Journal of information science*, 34(4):605–621, 2008. (Cited on pages 8, 21, 22, 24, and 99.)

[98] V-TFL. Team Fortress Classic. http://teamfortress.wikia.com/wiki/Team_Fortress_Classic. Accessed: 2013-11-11. (Cited on page 93.)

[99] Thomas W. Valente. Social network thresholds in the diffusion of innovations. *Social Networks*, 18(1):69–89, 1996. (Cited on pages 3 and 32.)

[100] Thomas W. Valente and Raquel Fosados. Diffusion of innovations and network segmentation: the part played by people in promoting health. *Sexually Transmitted Diseases*, 33(7):S23–S31, 2006. (Cited on pages 3 and 32.)

[101] Jacques Wainer, Michael Eckmann, Siome Goldenstein, and Anderson Rocha. How productivity and impact differ across computer science subareas. *Communications of the ACM*, 56(8):67–73, 2013. (Cited on pages 23, 25, and 99.)

[102] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*. Cambridge University Press, 2009. (Cited on pages 3, 12, 13, 17, 18, 19, and 26.)

[103] Steve Whittaker, Loen Terveen, Will Hill, and Lynn Cherny. The dynamics of mass interaction. In *Proceedings of the Conference on Computer Supported Cooperative Work and Social Computing (CSCW'98)*. ACM, 1998. (Cited on pages 30 and 31.)

[104] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. (Cited on pages 56 and 76.)

[105] Hao Wu, Jiajun Bu, Chun Chen, Can Wang, Guang Qiu, Lijun Zhang, and Jianfeng Shen. Modeling dynamic multi-topic discussions in online forums. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI '10)*. AAAI, 2010. (Cited on pages 31 and 69.)

[106] Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. Overlapping community detection in networks: the state of the art and comparative study. *ACM Computing Surveys*, 45(4), 2011. (Cited on pages 13, 14, 15, and 42.)

[107] Jaewon Yang and Jure Leskovec. Community-affiliation graph model for overlapping network community detection. In *International Conference on Data Mining (ICDM'12)*, pages 1170–1175. IEEE, 2012. (Cited on page 15.)

[108] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 3. ACM, 2012. (Cited on page 25.)

[109] Lofti A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications*, 9(1):149–184, 1983. (Cited on page 43.)

[110] Lotfi A. Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965. (Cited on pages 14, 115, and 116.)

[111] Ziming Zhuang, Ergin Elmacioglu, Dongwon Lee, and C. Lee Giles. Measuring conference quality by mining program committee characteristics. In *Proceedings of the ACM/IEEE-CS joint conference on Digital libraries*, pages 225–234. ACM, 2007. (Cited on page 27.)

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Final Version as of April 9, 2014 (`classicthesis` version 4.1).