Stability and Dynamics of Communities on Online Question-Answer Sites

Saskia Metzler[1]
Max-Planck-Institute for Informatics
Saarland Informatics Campus
Building E1.4
66123 Saarbrücken, Germany

Stephan Günnemann
Technical University of Munich
Department of Informatics & Institute for Advanced Study
Boltzmannstr. 3
85748 Garching, Germany

Pauli Miettinen[2]
University of Eastern Finland
School of Computing
P.O. box 1627
70211 Kuopio, Finland

12th December 2018

[1] Corresponding author. E-mail: saskia.metzler@mpi-inf.mpg.de

[2] Part of the work was done while the author was with the Max-Planck-Institute for Informatics.

Highlights

- We propose a hyperbolic structure for online question-answer forums.
- We show that the ratio of active members per community remains constant over time.
- This constancy shows irrespective of community size and in all datasets.
- This behaviour contrasts what is usually assumed to happen in online networks.

Stability and Dynamics of Communities on Online Question-Answer Sites

12th December 2018

Abstract

Social networks are present in our everyday lives, not just in face to face communication, but also when we communicate through the Internet. The latter leaves massive traces of data and thereby opens opportunities to acquire a better understanding of social communities in general. In this work, we are particularly interested in the patterns of volunteer efforts within the communities. To that end, we examine the community structure of several large online question-answer sites and how they evolve over time. To describe the user interaction patterns concisely, we employ the *hyperbolic community model.* This statistical model allows for a summary of each community in each time step by means of intuitive parameters that reflect the connectivity pattern within the network. Our study of the temporal evolution of these parameters reveals an important characteristic: In contrast to what has been observed earlier in the analyses of growth behaviour of online communities, we observe that the user activity within a community is constant with respect to its size throughout its lifetime. Furthermore, the structural organisation of different communities across different question-answer sites seems to follow a common scheme: There is a small group of users who is responsible for the majority of the social interactions.

Keywords: temporal network analysis | social network structure | hyperbolic community model | computational social science

1 Introduction

In this study, we investigate the particular dynamics of interactions between people in online communities on question-answer sites. Users of such sites donate their time and effort voluntarily to the community. In return, they gain visibility within the community through votes by other users. Besides the textual content that users provide, they leave traces of their interactions, such as who is responding to whom at which time.

We study freely available question-answer data from the large popular sites `reddit.com`, `stackexchange.com`, and `healthboards.com`. While `reddit.com` is a social news aggregation site with a very broad spectrum of topics, `stackexchange.com` is known for its free expert advice for the user asking a question, from which the entire community profits as well. Similarly, on `healthboards.com`, experts answer laymen's questions regarding health topics. Our primary result is the identification of a unifying pattern present in all examined groups: *The amount of active members is a constant fraction of the entire community throughout its lifetime.*

With our analysis, we add a large scale assessment of the volunteer effort in online social communities. Our results indicate that the active participation of only few community members within a group is a general organisational principle. Since online communication serves a social function [Wellman et al., 1996; Baym et al., 2004; Arnaboldi et al., 2013; Dunbar et al., 2015], this result is relevant for social communities in general.

While our analysis has a quantitative character, prior studies have analysed who these people are who volunteer for clubs, charities, or other organisations [Reed and Selbee, 2001; Nesbit and Gazley, 2012], and who are the contributors of Internet content [Bruns, 2008; Hargittai and Walejko, 2008]. Their motivations and backgrounds were the focus of these studies. In particular in the online world, contributions by people on a voluntary basis drive many communities. Wikipedia is one popular example of a community-driven project to which everybody can contribute. In contrast to the kind of user interactions we focus on, users of Wikipedia collaboratively edit the same document. They may even revert each others' changes. It has been shown that a large fraction of the content on Wikipedia is created by only a few highly active users [Matei and Bruno, 2015; Ortega et al., 2008], which is in accordance with observations of volunteer efforts in other contexts [Reed and

Selbee, 2001; Nesbit and Gazley, 2012]. While this inequality in participation is an important principle promotes helps productivity by forming leadership structures [Kittur and Kraut, 2008; Matei and Bruno, 2015], it might also be discriminative and demotivating for new contributors since it can be hard to compete with the well-established leadership core [Haklay, 2016]. It is exactly this competition for fame and personal visibility has also been found to be one of the central motivations for volunteering in online communities [Butler et al., 2007]. With our study, we identify the most active contributors of the analysed sites. We, however, gather no personal information about the users but rather study their interaction patterns on a large scale.

The dynamics within groups of individuals are a long-running research topic [Laumann and Pappi, 1976; Alba and Moore, 1978; Corradino, 1990; Morgan et al., 1997], but since only recently, with the increasing popularity of the Internet, large data sets can easily be acquired for analysis. Two decades ago researchers presented hypotheses about social communities after interviewing around 300 people multiple times within one year [Morgan et al., 1997], but such a dataset seems tiny nowadays. Today, millions of people leave traces of their interactions on the Internet. The availability of online network data has enabled numerous studies on the organisation of networks and the communities therein [Panzarasa et al., 2009; Leskovec et al., 2005, 2008; Cattuto, 2006; Ahn et al., 2007; Newman and Park, 2003; Sekara et al., 2016; Zhang et al., 2017].

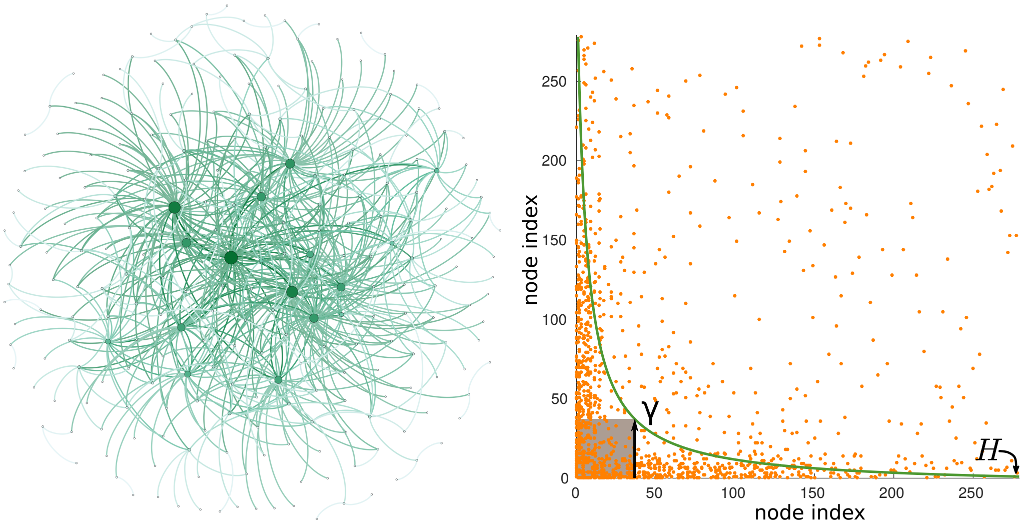
In our present work, we study the intra-community structure of various communities from three different online question-answer sites. We analyse which portion of the members is actively participating in discussions and how stable this core fraction is over an extended period of time. Especially in the early formation phase of a community, with new members joining month for month, our expectation was to see characteristic formation patterns. For example, Kumar et al. [2006], when studying the density of the network as a measure for the connectivity of every person in friendship networks, observed rapid growth followed by a decline and then slow but steady growth. In contrast, we find that for question-answer networks, the overall structure of a community remains very stable throughout its lifetime. A community might gain or lose active members, but relative to its size, the amount of core contributors remains nearly constant.

This observation raises the question as to *why* there is such a unifying organisational schema. We find that roughly 20 % of the members form the active core. This finding might hint towards deeper organisation principles. The 80–20 rule named after Pareto [Koch, 1998] resembles our result. This rule, however, does not allow for a hypotheses towards an underlying generative process to explain human social behaviour in general. Also, the extent to which this result is transferable to human social interaction in general remains an open question.

We believe that our modelling approach leads to a better understanding of network structures within social communities and thereby promotes the comprehension of the underlying processes of social interaction. In addition, we show how to concisely describe the intra-community structures on a high level.

2 The Hyperbolic Community Model

To describe communities in networks, we use the hyperbolic community model by Metzler et al. [2016]. This probabilistic model works on undirected, unweighted graphs and accounts for unequally strong ties among the community members. It is especially suitable to analyse the intra-community structure of social communities over time because it provides very intuitive parameters to summarise the connectivity pattern in every time step. It describes the shape of the degree-ordered adjacency matrix of an undirected graph by means of a specific function. This function is parametrised by two parameters: γ, indicating how large the core of a community is, and H, indicating how high the tail is (see Figure 1). As a special case, this model includes extreme structures such as a *star*, where one node is connected to everybody else but nobody has any further connection to each other, or a *clique*, where everybody is connected to each other. This flexibility enables the representation of a huge variety of communities. Our experiments in Sections 4.1.2 and 4.1.3 confirm that this model is a good match for the analysed data.



(a) Fruchterman–Reingold drawing of the user interaction graph, created with Gephi [Bastian et al., 2009].

(b) Fitted model of the commmunity on the degree-ordered adjacency matrix with the core area shaded in grey. Dots indicate edges between nodes i and j.

Figure 1: The gardening community from Stackexchange in June 2016 and its fitted model. The community constitutes 279 users, according to the fitted hyperbolic model, $\gamma = 38$ are core members, and the tail H thins down to two members.

2.1 Formal Definition

Following Metzler et al. [2016], for an undirected graph $G = (V, E)$ with n nodes and m edges, a number from $\{0, \dots, n-1\}$ is assigned to every vertex, and (i, j) is used to denote both a pair of vertices and the (potential) undirected edge between them. The graph will be represented using its *adjacency matrix* $\mathbf{A} = (a_{ij}) \in \{0, 1\}^{n \times n}$.

A *community* C is a tuple (V_C, π_C, Θ_C). The set $V_C \subseteq V$ contains the nodes in the community, and we write $n_C = |V_C|$. The permutation $\pi \colon V_C \to \{0, \dots, n_C - 1\}$ orders the nodes. In general, we assume the nodes to be ordered according to their degrees inside the community. Θ denotes the set of parameters needed to describe the model.

The crucial part is that not every edge between the nodes in V_C is necessarily part of the inside-community area. The community model is defined using a binary decision function

$$f\colon \{0,\dots,n_C-1\}\times\{0,\dots,n_C-1\}\times\Theta\to\{0,1\}$$

operating on a parameter set Θ and deciding for any pair of vertices $(i,j)\in\{0,\dots,n_C-1\}\times\{0,\dots,n_C-1\}$ if an edge between i and j belongs to the inside-community area. Only in the extreme case where f decides positive for every edge (i,j), the community resembles a clique.

The decision function used in the hyperbolic community model is the *hyperbolic function*

$$f\bigl(i,j,(p,\tau)\bigr)=(i+p)(j+p)\le\tau$$

which requires the parameter set $\Theta=\{p,\tau\}$. This function yields a hyperbolically shaped decision boundary in the degree-ordered adjacency matrix (as depicted in Figure 1b). For easier interpretability of the model, we use the equivalent re-parametrisation of the model into $\Theta=\{H,\gamma\}$. These parameters define two points of the hyperbolic curve: the point at which it crosses the diagonal (i.e. when $i=j$), and the point at which the hyperbola exits the community (i.e. j for which $i=n_C$ or vice versa), completely defining the shape of the hyperbola (see Figure 1b). Parameters γ and H can, respectively, be interpreted as the *size of the core*, that is, the densely connected (quasi-) clique of the community, and as the *thickness of the tail*, that is, how connected the least-connected members of the community are.

Since real-world communities are rarely ever exact, we apply a probabilistic estimate. We assume that edges $(i,j)\in V_C\times V_C$ are drawn from a Bernoulli distribution, $a_{i,j}\sim\text{Bernoulli}(p_*)$, where $\mathbf{A}=(a_{ij})$ is the adjacency matrix of the graph and p_* the *density* of the area that the edge belongs to. For a community, we have two kinds of areas: the inside-community area A_C and its complement $\overline{A_C}$. We denote their respective densities by

$$d_C=|E_C|\,/\,|A_C| \tag{1}$$

and

$$d_O=|E\cap\overline{A_C}|/\left|\overline{A_C}\right|\ . \tag{2}$$

These densities correspond to the maximum-likelihood solutions of the variables p_* for the edges that are inside or outside of the community area. Let us write $G|_{V_C}$ to denote the subgraph of G induced by the nodes of the community C. To find the best model we consider the likelihood of the subgraph $G|_{V_C}$ given community C, $L(G|_{V_C} \mid C)$. We compute

$$\begin{aligned}\log L(G|_{V_C} \mid C) = {} & |E_C| \log(d_C) + |\overline{E_C}| \log(1 - d_C) \\ & + |E \cap \overline{A_C}| \log(d_O) + |\overline{A_C} \setminus E| \log(1 - d_O)\end{aligned}$$

for all permissible combinations of γ and H and keep the best (see Metzler et al. [2016] for explanations as to why this exhaustive search is feasible).

2.2 Related Approaches

The widely recognised core-periphery model [Borgatti and Everett, 1999] can be regarded as a special case of the hyperbolic community model. In its extreme, it assumes the core to be completely connected and all nodes in the periphery to be connected to all nodes from the core but not to each other. The employed model [Metzler et al., 2016] is conceptually related but has more freedom when it comes to the shape of the community: It accounts for the fact that not all nodes in the periphery are equally well connected to the core. Furthermore, it can deal with imperfect data.

Communities in real networks typically do not show a uniform density profile [Araujo et al., 2014]. Hence, many alternative models used in community detection approaches are not suited for our means. Despite their diversity, they mostly assume a community to be a *block-shaped* area with *uniform density* in the adjacency matrix. This quasi-clique assumption is the basis of prominent techniques such as stochastic block-models [Nowicki and Snijders, 2001; Airoldi et al., 2008], affiliation network models [Yang and Leskovec, 2012], pattern based techniques such as the detection of quasi-cliques [Günnemann et al., 2014; Boden et al., 2012], and cross-associations [Chakrabarti et al., 2004].

An alternative approach would be to think of the structure of networks as a hierarchy where small sub-communities compose into larger ones [Girvan and Newman, 2002; Palla et al., 2005; Lancichinetti et al., 2009]. Yang and Leskovec [2012] explain non-uniform density inside communities as the result of overlap between communities considering edges to be more likely due to the combined density of the overlapping tiles. Here, hierarchical methods appear inappropriate, mainly for the lack of rich annotations in the data sets. Also, we observe a clearly non-uniform distribution of the edges in (often truly) non-overlapping communities, invalidating the prior assumption of Yang and Leskovec [2012].

Building on the model of Borgatti and Everett [1999], another idea has been proposed by Rombach et al. [2014]: Nodes of a network are assigned *core scores* along a continuous spectrum that reflect how deep a node lies inside the core. This measure can also be used for a discrete decision of whether a node belongs to the core. With an appropriately chosen threshold, we would expect generally comparable results for our study using this model. An interesting alternative direction of research is to explore the exact correspondence of this approach to that of Metzler et al. [2016]. For the present study, we focus on analysing the different question-answer sites under the assumption of the latter model. Metzler et al. [2016] state good scalability of the method while Rombach et al. [2014] report insufficient performance; the largest examined network had 552 nodes, which is orders of magnitude smaller than the data we analyse here. Apart from the need for a scalable approach, an advantage of employing the Rombach et al. [2014] model could be that it allows for weighted networks and, therefore, could incorporate more information about the strengths of user interactions. The approach of Metzler et al. [2016] would need to be extended to consider weighted graphs. See Section 5.2 for a discussion about the benefits of dichotomous data.

3 Datasets

We examine meta-information of three large online discussion sites. We are interested in identifying the active users within the different communities of these sites. Therefore we collect the user interactions, that is, who has replied to whose post at which time. This information reveals the interactions between users and allows for the construction of user interaction graphs for each data set. Every obtained graph denotes the users as nodes and the interactions, labelled with times of occurrence, as edges. A summary of the employed data sets and their characteristics are displayed in Table 1. All data is publicly available. Further details on the data preparation are provided in Section S1 in the Supplementary Material.

Table 1: Characteristics of the datasets. The statistics about the community size refer to the number of communities considered in this study and are reported with respect to their number of nodes. We require communities to have more than 100 nodes in total and to cover a time span of more than 12 months.

	Reddit	Stackexchange		HealthBoards
		forums	meta-forums	
covered time span	2005–2016	2008–2016		1999–2013
nodes	$\sim 230\times 10^6$	$\sim 8\times 10^6$		338 079
communities	635 048	160	160	235
used communities	6 056	147	117	222
max. community size	155 511	219 693	30 223	18 924
avg. community size	2 774	15 124	849	3 015
avg. time span (months)	42	58	58	128

3.1 Reddit

The largest of the analysed sites is `reddit.com`. Reddit is an American social news aggregation and discussion website. As of the end of 2016, it encompasses 635 048 different topics, called *subreddits*. In every subreddit, users can open new discussion threads where other users can comment. From these threads, we can gather who answered whom and when and construct a labelled undirected graph. Users constitute the nodes of this graph. There is an edge between two nodes if a user replied to another user (ignoring self-edges). Every edge is labelled with the date of the interaction. Furthermore, we group the nodes into communities by the subreddit under which the users contributed something.

Notice that we do not analyse whether the initiator of a discussion thread actually poses a question. While Stackexchange and HealthBoards have a clear question-answer structure, Reddit is a discussion board and may well have threads where no question is posed or answered. This difference in the type of content, in addition to its size, makes the dataset particularly interesting for the present study: The fact that a post encourages other users to reply and engage in a discussion allows us to analyse this dataset in the same way as we do the pure question-answer sites. At the same time, we might expect to observe differences in the results due to the difference in the type of content.

3.2 Stackexchange

We likewise obtained an undirected labelled graph from `stackexchange.com`. Stackexchange is composed of question-answer websites on topics in various fields, the most prominent being related to computer programming and system administration. As of the end of 2016, there are a total of 160 different topics, each of them with a meta-discussion board. The derived graph contains the users as nodes. There is an edge between two nodes if a user replied to a question of another user or if a user commented on a post (can be question or answer) of another user. Every edge is labelled with the timestamp of the interaction. In addition, to define the communities, we group the nodes with respect to the topic under which they made a contribution.

3.3 HealthBoards

The web page `healthboards.com` is a long-running message board for patient to patient health support. It consists of 235 message boards for different health-related topics. This data is orders of magnitude smaller than the aforementioned resources. However, it is particularly interesting because not only the formation of communities can be observed, but also the dissolving phase where user activity gradually declines. The graph we derive has the users as nodes. Edges are formed through every answer of one user to a thread opened by another user and are annotated with the timestamp of the interaction. The nodes are grouped in communities according to the message boards where they posted.

4 Results

Online social communities reflect social relationships between people [Wellman et al., 1996; Baym et al., 2004]. Understanding the communication patterns within large online discussion boards might thus enable a better understanding of the dynamics within social networks in general. To that end, we collect the information as to who responded to whom at which time from the different forums of `reddit.com`, `stackexchange.com`, and `healthboards.com`. We regard each different forum, such as `bitcoin.stackexchange.com`, as a community and study the development of the connectivity patterns within each of them over monthly time steps.

4.1 Model Quality

We start with a discussion of modelling decisions for representing the data and an assessment of the suitability of the chosen model.

4.1.1 Modelling Decisions

We use monthly intervals to discretise the time line. This choice trades off between a fine-grained view on the evolution of the communities and keeping the amount of models to compute and evaluate within a feasible range.

We consider the subgraph of every community individually instead of modelling the whole graph at once, as (1) we want to focus only on the intra-community interactions, (2) for many communities there is very little overlap, and (3) the computational complexity would get unreasonably large compared to the additional information gain.

We experimented with how to accumulate the data for the monthly time steps. Options are for every time step (1) to accumulate all interactions from the beginning of the time series to the current time step, (2) to use a sliding window in order to accumulate over the last few months until the current time step, or (3) to only take the interactions of the respective month as edges. The results presented here were obtained employing the latter option, which exhibits the highest variance and, thus, is the hardest case for what we aim to show. The other two options yield denser subgraphs, and because there is overlap between the data of consecutive time steps, the course of the core size parameter γ is much smoother. As discussed in detail in Section S3, we obtain similar results with respect to the relative shapes of the communities with these set-ups.

We report only results for communities with at least 100 nodes in total and a covered time range of more than 12 months.

4.1.2 Comparison to Block Models

To validate that the hyperbolic community model is a good way to describe the analysed data sets, we compare it to the commonly used alternative, a quasi-clique model.

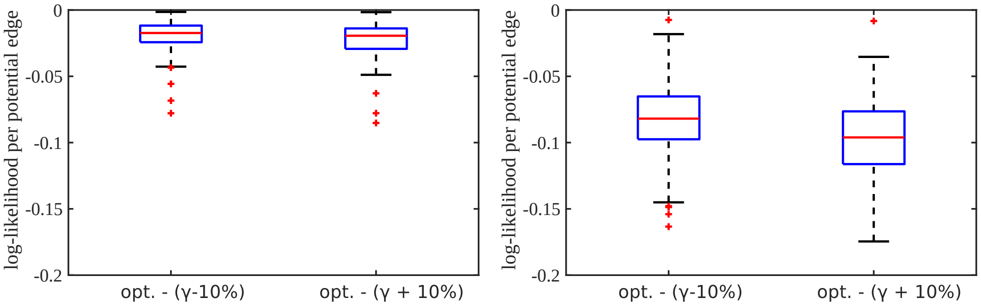
A quasi-clique, or in other words, a uniformly dense block, is a common description of a community in a graph. The hyperbolic community model includes this option as a special case, i.e., when the core size γ is equal to the size of the community. We want to validate that such a model is not capable of describing the data as well as the hyperbolic community model. We use the log-likelihood to judge the description quality.

From every community, from every dataset, we take the timestep where the community is the largest. This usually coincides with the last timestep. Only in HealthBoards does user activity decrease towards the end of every community's time series. Very small communities are susceptible to noise and are therefore not the focus of this analysis.

We carry out a likelihood ratio test. The null hypothesis H_0 is that all parameters are fixed so that they create block structures. The alternative hypothesis H_1 is that the parameters are not fixed. The likelihood ratio test statistics are given by $\lambda = 2\log\left(L(\text{hyperbolic model})/L(\text{block model})\right)$. We find that all Stackexchange communities are statistically significantly better explained by the hyperbolic block model (significance level $\alpha = 0.01$). For Reddit, 99.3 % of the communities are better explained with the hyperbolic block model at $\alpha = 0.01$, and 99.7 % at $\alpha = 0.05$. For HealthBoards, 94.1 % of the communities are better explained at $\alpha = 0.01$, and 95.5 % at $\alpha = 0.05$. It should be noted that the cases where a hyperbolic community model is not statistically significantly better than the block model coincide with extremely small communities, typically below twenty nodes at their maximum.

4.1.3 Robustness of the Models

We analyse the robustness of the obtained hyperbolic community models. To that end, we analyse how the log-likelihood deviates if we alter the optimal value for the parameter γ by 10 % (relative to number of active members) in each direction. Notice that the absolute log-likelihood of a community model is dependent on the size of the community. To compare among different

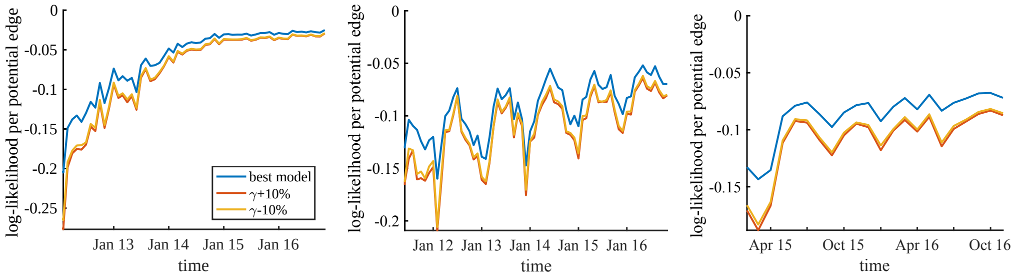


(a) Non-meta communities of Stackexchange (b) Meta-communities of Stackexchange

Figure 2: Differences in the log-likelihood per possible edge when applying the best hyperbolic community model versus shifting the parameter γ by 10 % away from its optimum. The distributions are the average log-likelihood differences per community. The boxplots display distributions such that in each box, the central mark is the median and the edges of the box are the first and third quartiles. The whiskers extend to the most extreme data points that are not outliers. Points are considered outliers if they are larger (smaller) than the third (first) quantile plus (minus) 1.5 times the difference between the quantiles. Outliers are plotted individually.

communities, we normalise the log-likelihood by the number of possible edges within the community, i.e., $(N^2 - N)/2$ for a community with N members.

We observe that, indeed, the log-likelihood never improves when the parameter is shifted away from the optimum. More importantly, the log-likelihood worsens by a similar amount in every time step and in every community and for both directions of shifting γs (see Figure 2). When increasing γ by 10 %, we notice a drop of 0.022 on average for the non-meta Stackexchange communities and 0.096 for the meta communities. Likewise, when decreasing γ by 10 %, the average drop is 0.019 and 0.083, respectively. To illustrate the development of the log-likelihood with the altered parameter γ over time, exemplary courses of the evolution are displayed in Figure 3. We conclude that the hyperbolic community model is very robust and thus well-suited for our analysis.



(a) Academia (b) Gardening (c) VI

Figure 3: Evolution of the log-likelihood of the hyperbolic community model for selected Stackexchange communities (blue) and the log-likelihood when shifting the parameter γ 10 % away from its optimum (red and yellow). The log-likelihood is reported relative to the number of possible edges within the community.

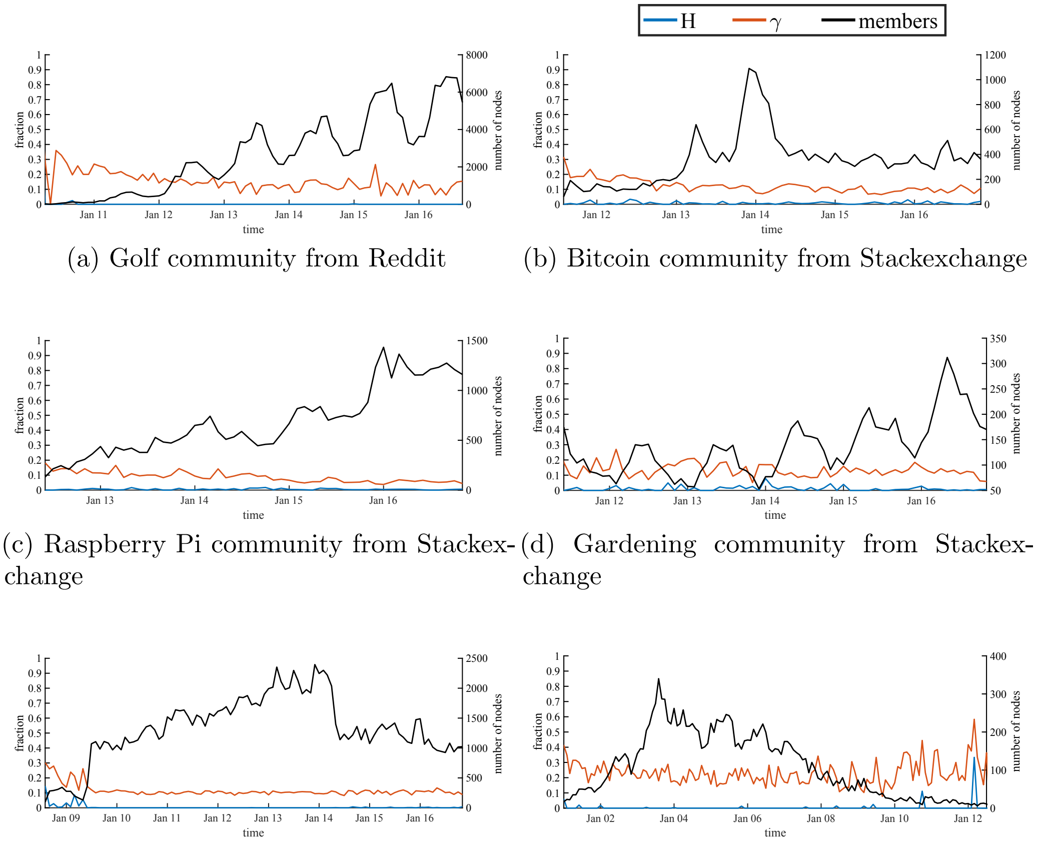
4.2 Fitting Community Models

We fitted a hyperbolic community model via log-likelihood optimisation for every time step and for every community of the datasets.[1] A collection of examples of the behaviour of the model parameters over time is displayed in Figure 4.[2] We observe a predominant shape of the intra-community connectivity pattern: Between 10 % and 30 % of nodes form the core. The rest of the nodes are loosely connected to the core area. As can be observed in Figure 4 and is examined in more detail subsequently, the value of γ is almost constant in this range over the full lifetime of the community. Surprisingly, we see this shape throughout the whole lifetime of every (sufficiently large) community, suggesting that the core of each community is invariant to its size and life time. Of course some fluctuations occur if there are very few community members in total, such as in the initiation phase of a community.

To quantify the characteristics of the observed shapes, we examine parameter distributions for each dataset. As summarised in Figure 5, the distributions for the core size parameter γ predominantly range between 10 % and 30 %, while the tail H is mostly very thin. The cores of the Stackexchange

[1] The code for the hyperbolic community model is publicly available [Metzler et al., 2016]. All computations have been carried out using Matlab.

[2] All time line plots and video presentations of selected models can be found in the Supplementary Material, Sections S8 and S9.



(a) Golf community from Reddit

(b) Bitcoin community from Stackexchange

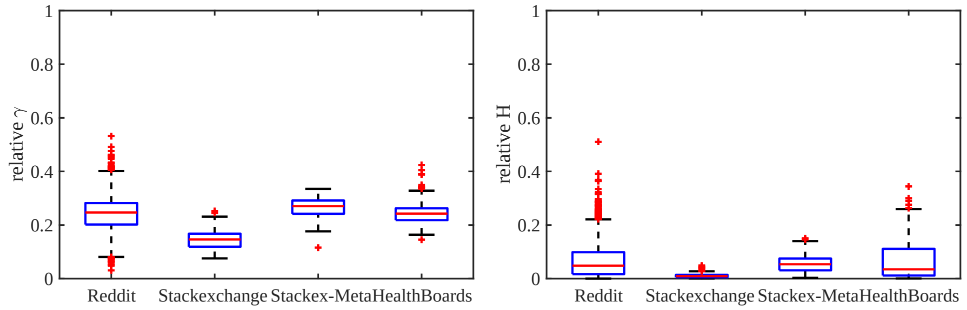
(c) Raspberry Pi community from Stackexchange

(d) Gardening community from Stackexchange

(e) Meta-stackexchange community from Stackexchange

(f) Beauty & Cosmetics community from HealthBoards

Figure 4: Examples of models for communities from the different question-answer sites over time. The model parameters H and γ obtained from fitting the hyperbolic community model are shown relative to the community size. The right axis in every plot denotes the absolute size of the community (black curve). Only data from communities with at least 20 members is displayed. Further time lines are displayed in Figures S9 to S16.



(a) Distribution of the core size γ (b) Distribution of tail height H

Figure 5: Distribution of the model parameters H and γ. Each box displays the distribution of averages over the time lines of each community in the respective data set. Further statistics are displayed in Table S2.

communities are even smaller, with their median size being 15 %. We regard the actual discussion boards of Stackexchange and the meta-communities separately, where discussions about the respective board take place. Meta-communities are much smaller, their activities often vary heavily, and their members are a subset of the users of the respective basic discussion board. We find the cores of these communities to be 26 % of the community size on average, which is not even within the 90 th percentile of the basic Stackexchange communities. This is likely a result of the bias towards otherwise active users as participants in these discussions.

4.3 Regression with Respect to Time

We analyse whether the seemingly constant relative size of the core of the communities is truly constant over time. To that end, we perform two tests. First, we fit a linear model on each time line of γs. To assert that no quality is gained when using more complex models, we compare the fitting quality to that of higher order polynomials in Section S4. As Figure 6a shows, we observe that the slope of the linear models is close to zero everywhere. Mostly large p-values (see Table S2) provide no evidence to reject the null-hypothesis, which states that the slope is equal to zero. Since this alone is no proof of the significance of our finding, we secondly measure the Pearson correlation coefficient between each time line of γ and a constant function. We find a strong correlation throughout the data with high significance.

This statement may seem incoherent with the definition of Pearson's correlation coefficient: For a sample statistic, the correlation r between variables x and y is

$$r = \frac{\sum\limits_{i=1}^{n}(x_i - x)(y_i - y)}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2 \sum\limits_{i=1}^{n}(y_i - \bar{y})^2}}$$

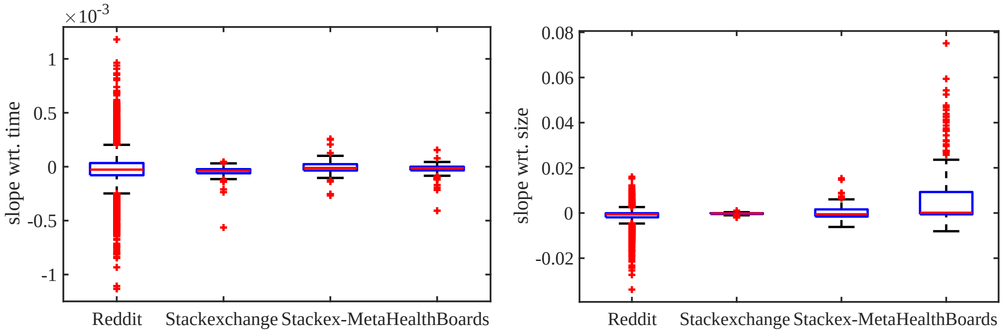
where x_i and y_i are the values of x and y for the ith sample. In the case we describe, all y_i would be identical and in particular $y_i = \bar{y}$ which yields 0 in the denominator. Commonly, $r = 0$ is the defined outcome for this ill-defined case. To remedy this deficiency in the definition, we tilt the coordinate system prior to computing the correlation coefficient. To that end, we apply a coordinate transformation to all samples by multiplication with the rotation matrix

$$m = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \qquad \text{where} \qquad \alpha = \frac{\phi \cdot 2\pi}{360°}.$$

This way, we achieve a turning of the data by an arbitrarily chosen angle of $\phi = 45°$ and may compute the Pearson correlation coefficient in this transformed space. The argument for this procedure is that if there is no correlation and the data is just normally distributed along both axes, then no rotation will reveal any correlation. Vice versa, however, if there is any correlation with the flat line, then we can see it by rotation.

The p-values of the correlation are computed for testing the hypothesis of no correlation against the alternative that there is a non-zero correlation. For every data set, we observe a significant amount of correlation, with no p-value larger than 0.000001 (see Table S2).

In particular, we notice that all examined datasets of the different question-answer boards show a similarly constant core size. While Reddit and Stack-exchange data range over less than five years on average and include mostly growing communities, the average HealthBoards communities have user interactions recorded over more than a decade. Even within such extended time periods, which include the formation and dissolving of communities, we have identified the constant core size as a unifying pattern.



(a) Time as x-axis.

(b) Community size as x-axis.

Figure 6: Distribution of the slopes obtained from linear regression on every community.

4.4 Regression with Respect to Community Size

We furthermore study whether there is also no trend in the core size with respect to the size of a community. With this experiment, we aim to confirm what is observable in Figure 4: Over time, the community sizes vary, but the sizes of community cores are invariant to community size. To that end, we compute regression models for every community, like we did in the previous section, this time, however, with the community size instead of time on the x-axis. As Figure 6b shows, the relative core size is close to constant, also under variation of community size. We can therefore assert that the size of the cores is invariant to changes in the community size. Like for the regression with respect to time, we measure the Pearson correlation coefficient towards a constant function also in this setup. We again observe a strong correlation throughout the data with high significance (all p-values < 0.000001, compare with Table S2).

As Figure 6b indicates, outliers are an order of magnitude more extreme in this set up and appear to be more prominent, especially in the HealthBoards data. A closer look at the data (see Section S7 and, in particular, Figure S7) reveals that the smallest of the communities constitute the outliers in these boxplots and are most likely explained by insufficient data to compute reliable hyperbolic models.

4.5 Stability of the Core

Our results above indicate that the relative *size* of the cores stays constant; however, we do not see whether the involved users inside the core vary strongly over time. To analyse how constant the fraction of core members stays from one time step to the next, we measured the stability of the core using the containment index. Given two sets of nodes, this index indicates which percentage of nodes from the smaller set is contained in both sets. Formally, given two sets of nodes, $\mathcal{G}$ and $\mathcal{H}$, the containment index $C(\mathcal{G}, \mathcal{H})$ is defined as

$$C(\mathcal{G}, \mathcal{H}) = \frac{|\mathcal{G} \cup \mathcal{H}|}{\min\left(|\mathcal{G}|, |\mathcal{H}|\right)} .$$

We observe that from one time step to the next, between 40 % and 60 % of the nodes are overlapping. Figure 7 displays the result per dataset and includes the comparison of cores that are 2, 4, and 8 time steps apart. As expected, we observe a gradual decrease of the overlap. However, even between time intervals that are more than half a year apart, the overlap is more than 30 % for all datasets. This suggests that a substantial portion of the active members stay active for a longer period of time and might even be a conservative estimate, since the employed measure ignores users who take a pause and then return.

We further observe that the cores of the Stackexchange communities and, in particular, of the meta-discussion boards are more stable than those of Reddit or HealthBoards. This suggests that active Stackexchange users generally stay more committed to their community, and, in particular, those users who discuss their community on a meta-level adhere to their communities. Reddit users, on the other hand, are less devoted to their communities, called *subreddits*. This is logical, as new subreddits can be opened easily, and many of them are about arbitrary topics relating to current events.

For both HealthBoards and Stackexchange, we observe a stronger decline in the fraction of steady members when increasing the time interval than we do for Reddit or the meta-forums of Stackexchange. We hypothesise that this is due to the expert-layman interaction nature in these two boards: A layman comes, discusses his matters with the experts, and leaves again.

An interesting direction of future research is to characterise the active users who constitute the core, and especially those who stabilise the community

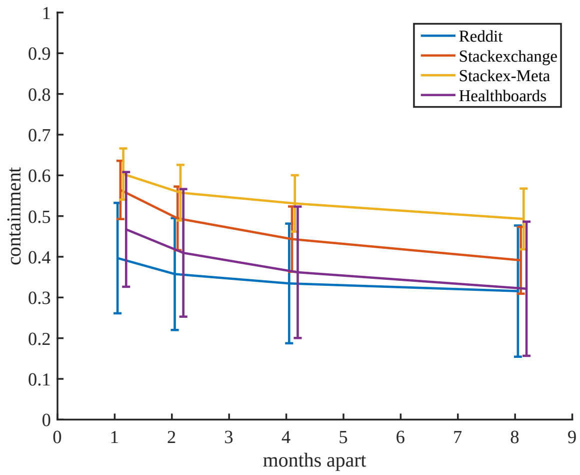


Figure 7: Core overlap between time-wise subsequent cores, and cores 2, 3, and 4 months apart. The exact numbers can be found in Table S2.

over time. For Stackexchange, where each user has a *reputation* score that reflects expertise, we analyse to what extent high reputation is correlated to users in the cores of communities. As we detail in Section S5, high reputation alone is, however, not a reliable indicator of the activity of a user.

5 Discussion and Conclusions

Modelling the interaction graphs of various communities of different online question-answer sites, we observe very similar patterns that remain stable over the lifetime of every community. We find that a constantly small fraction of the members constitutes the active core of each community, donating their time and effort voluntarily to the entire group.

5.1 Placement with Respect to Earlier Work

One might think that particular communities have a substantially higher fraction of active members than others. Or, communities in the early phase, during their formation, might initially start as a block-like structure, where everybody interacts with everybody else. This phase might then be followed by a star pattern, as new members join and connect first to the old members. Such a formation pattern has been reported by Kumar et al. [2006] for the social networks Flickr and Yahoo! 360, but cannot be confirmed for the analysed data sets.

Arguably, it is a matter of conjecture whether differences in the data analysis disguise such patterns or whether they are truly absent. Kumar et al. [2006] and Leskovec et al. [2005] base their findings on the density observed in whole networks, whereas we study the hyperbolic community model. The major difference is that we consider only those nodes that have interactions with others while Kumar et al. [2006] consider every node irrespective of the connectivity. It is likely that the pattern they reported is caused by singleton users signing up and not immediately building up connections. Furthermore, Kumar et al. use discretisation based on weeks' granularity; we use months. However, since their detected patterns occur within a period of 40 weeks, they should also be visible when studying monthly patterns.

Another important distinction between our study and earlier work is that in previous representations of social networks [e.g. Leskovec et al., 2005; Kumar et al., 2006], an edge corresponds to a *passive* friendship relation. That means, adding more friends to someone's network does not, per se, increase the workload of this person. This is in contrast to our setting, where each edge represents *active* participation in the question-answer site, and especially in case of complex answers, can involve significant time investment. Such active commitment, done at least partially for altruistic reasons [Nesbit and Gazley, 2012; Butler et al., 2007], is likely to follow different dynamics than the mostly passive inclusion of friends.

One possible organisational principle underlying the networks studied here is the widely applied Pareto principle, or the *law of the vital few* [Koch, 1998], which states that around 80 % of the effects come from around 20 % of the causes. This principle matches observations in various fields, such as business and biology [Fernández-Sánchez and Rodríguez-López, 2010; Daly and Farley, 2011; Jankowski et al., 2013] and also explains communication in online discussion boards surprisingly well: about 20 % of the people are responsible for 80 % of the content in every community. The actual people who contribute actively might change over time, but the structural stability appears to be a property of large groups. Palla et al. [2007] emphasise that small groups, on the contrary, are susceptible to change and are likely to dissolve if their members change. This observation agrees with the increased amounts of variance we observe within small communities.

From the viewpoint of an individual, preceding studies [Roberts et al., 2009; Arnaboldi et al., 2013; Dunbar, 1992; Dunbar et al., 2015; De Salve et al., 2016] yield supporting evidence for more general organisation principles in human communication. Saramäki et al. [2014] coin the term *social signature*, suggesting that most communication between individuals is limited to a small portion thereof. The frequent contacts of a person might change over time, but the overall distribution persists. The authors hypothesise that this is a consequence of finite resources, such as time available for communication and emotional capital. Following this thought, most active users in the examined question-answer sites would either have extremely frequent contact with their closest friends, or they have no further social circle. As either scenario seems unlikely, a subsequent qualitative analysis might be necessary to unify these results. Our study covers voluntary effort in the form of question answering. We assume that this serves as a proxy to communication behaviour in general. However, the conducted study is limited to the communication that people "donate" to the examined sites. The large amount of studied data allows for statistically well-grounded hypotheses, but we cannot make holistic claims about the communication patterns from the perspective of an individual.

5.2 Data Representation

To construct the studied interaction graphs, we use the response of one user to the posting of another as an indicator for a social interaction. For simplicity, these interactions are modelled dichotomously and we do not account for the direction of the interactions. Incorporating the latter would require a more complex community model in which donators and receivers are distinguishable. What further insights could be gained from such a modelling remains an open question. In the current approach, we seek for a very simple summary of the communities under study, describing their shape by just two different parameters. This allows us to summarise especially large data sets into a graspable format.

To model additional information about the strength of the edges, every interaction could be weighted, either by the amount of interactions in a given time interval, or by the elapsed time between a post and its response. Empirical examination of Stackexchange communities indicates that the majority of connections, in the former setting, would show very low weights, mostly equal to 1 and, therefore, would not provide much additional information (see Section S2). The reason is potentially the particular expert-layman interaction on this site. Whether more general discussion boards like on Reddit could benefit more from weighted interaction graph models, and how the time between interactions as an indicator of strength could be beneficial, are interesting directions of future research. Such an analysis would also require the use of a modified approach to model the communities. The model by Metzler et al. [2016] in its current formulation is restricted to undirected, unweighted networks.

5.3 Use for Finding Discontinuities

We have seen that the relative amount of active members in a community remains constant over time. In particular, this holds while the actual size of the community might vary—either through growth in general or due to seasonal trends (such as in the gardening community of Stackexchange, or the golf community of Reddit (Figures 4a and 4d). Occasionally, however, seemingly independent of the community size, we observe short-term enlargements of the core size by about 10 % (Figure 4c). The Raspberry Pi community of Stackexchange, for instance, shows such discontinuities around March 2013 and April 2014. Coincidentally, these dates refer to release dates of new versions of the product. Unlike for the Bitcoin community (Figure 4b), where the size of the community can easily be aligned with major events in the value development of the currency, the Raspberry Pi release events do not show in the number of members, rather in their connectivity patterns. Our employed modelling framework might thus facilitate new methods for event detection.

5.4 Summary

In summary, our work provides new insight on the evolution of community structures in large networks. The examined online question-answer sites show a common scheme of roughly 20 % of highly connected active members and 80 % loosely associated members who mainly communicate with the active core. This scheme remains throughout the lifetime of a community.

References

Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In C. Williamson, M. E. Zurko, P. Patel-Schneider, and P. Shenoy, editors, *Proceedings of the 16th international conference on World Wide Web (WWW)*, pages 835–844, New York, NY, USA, 2007. ACM.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *J Mach Learn Res*, 9:1981–2014, 2008.

R. D. Alba and G. Moore. Elite social circles. *Sociol Methods Res*, 7(2): 167–188, 1978.

M. Araujo, S. Günnemann, G. Mateos, and C. Faloutsos. Beyond blocks: Hyperbolic community detection. In T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, editors, *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECMLPKDD)*, pages 50–65, Heidelberg, Germany, 2014. Springer.

V. Arnaboldi, A. Guazzini, and A. Passarella. Egocentric online social networks: Analysis of key features and prediction of tie strength in Facebook. *Comput Commun*, 36(10-11):1130–1144, 2013.

M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In E. Adar, M. Hurst, T. Finin, N. Glance, N. Nicolov, and B. Tseng, editors, *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM)*, Menlo Park, California, USA, 2009. The AAAI Press.

N. K. Baym, Y. B. Zhang, and M.-C. Lin. Social interactions across media. *New Media Soc*, 6(3):299–318, 2004.

B. Boden, S. Günnemann, H. Hoffmann, and T. Seidl. Mining coherent subgraphs in multi-layer graphs with edge labels. In Q. Yang, D. Agarwal, and J. Pei, editors, *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1258–1266, New York, NY, USA, 2012. ACM.

S. P. Borgatti and M. G. Everett. Models of core/periphery structures. *Soc Networks*, 21:375–395, 1999.

A. Bruns. *Blogs, Wikipedia, Second Life, and Beyond.* Peter Lang, Bern, Switzerland, 2008.

B. Butler, L. Sproull, S. Kiesler, and R. Kraut. Community Effort in Online Groups: Who Does the Work and Why? In S. P. Weisband, editor, *Leadership at a Distance*, chapter 9, pages 171–194. Psychology Press, New York, USA, 2007.

C. Cattuto. Semiotic dynamics in online social communities. *Eur Phys J C Part Fields*, 37:33–37, 2006.

D. Chakrabarti, S. Papadimitriou, D. S. Modha, and C. Faloutsos. Fully automatic cross-associations. In W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel, editors, *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 79–88, New York, NY, USA, 2004. ACM.

C. Corradino. Proximity structure in a captive colony of Japanese monkeys (Macaca fuscata fuscata): An application of multidimensional scaling. *Primates*, 31(3):351–362, 1990.

H. E. Daly and J. Farley. *Ecological economics: Principles and Applications.* Island press, 2011.

A. De Salve, M. Dondio, B. Guidi, and L. Ricci. The impact of user's availability on On-line Ego Networks: A Facebook analysis. *Comput Commun*, 73:211–218, 2016.

R. I. M. Dunbar. Neocortex size as a constraint on group size in primates. *J Hum Evol*, 22(6):469–493, 1992.

R. I. M. Dunbar, V. Arnaboldi, M. Conti, and A. Passarella. The structure of online social networks mirrors those in the offline world. *Soc Networks*, 43:39–47, 2015.

G. Fernández-Sánchez and F. Rodríguez-López. A methodology to identify sustainability indicators in construction project management—application to infrastructure projects in spain. *Ecol Indic*, 10(6):1193–1201, 2010.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–7826, 2002.

S. Günnemann, I. Färber, B. Boden, and T. Seidl. GAMer: a synthesis of subspace clustering and dense subgraph mining. *Knowl Inf Syst*, 40(2): 243–278, 2014.

M. Haklay. Why is participation inequality important? In C. Capineri, M. Haklay, H. Huang, V. Antoniou, J. Kettunen, F. Ostermann, and R. Purves, editors, *European Handbook of Crowdsourced Geographic Information*, pages 35–44. Ubiquity Press, London, UK, 2016.

E. Hargittai and G. Walejko. The Participation Divide: Content creation in the digital age. *Inf Commun Soc*, 11(2):239–256, 2008.

M. D. Jankowski, C. J. Williams, J. M. Fair, and J. C. Owen. Birds shed RNA-viruses according to the Pareto principle. *PLoS One*, 8(8):1–9, 2013.

A. Kittur and R. E. Kraut. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In B. Begole and D. W. McDonald, editors, *Proceedings of the ACM 2008 conference on Computer supported cooperative work*, pages 37–46, New York, NY, USA, 2008. ACM.

R. Koch. *The 80/20 Principle: The Secret of Achieving More with Less.* A Currency book. Doubleday, New York, USA, 1998.

R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In L. Ungar, M. Craven, D. Gunopulos, and T. Eliassi-Rad, editors, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 611–617, New York, NY, USA, 2006. ACM.

A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New J Phys*, 11(3):033015, 2009.

E. O. Laumann and F. U. Pappi. *Networks of Collective Action: A Perspective on Community Influence Systems.* Academic Press, New York, USA, 1976.

J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In R. Grossman, R. Bayardo, and K. Bennett, editors, *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 177–187, New York, NY, USA, 2005. ACM.

J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, editors, *Proceedings of the 17th international conference on World Wide Web (WWW)*, pages 695–704, New York, NY, USA, 2008. ACM.

S. A. Matei and R. J. Bruno. Pareto's 80/20 law and social differentiation: A social entropy perspective. *Public Relat Rev*, 41(2):178–186, 2015.

S. Metzler, S. Günnemann, and P. Miettinen. Hyperbolae are no hyperbole: Modelling communities that are not cliques. In F. Bonchi, J. Domingo-Ferrer, R. Baeza-Yates, Z.-H. Zhou, and X. Wu, editors, *IEEE 16th International Conference on Data Mining (ICDM)*, pages 330–339, Los Alamitos, California, USA, 2016. IEEE Computer Society.

D. L. Morgan, M. B. Neal, and P. Carder. The stability of core and peripheral networks over time. *Soc Networks*, 19(1):9–25, 1997.

R. Nesbit and B. Gazley. Patterns of Volunteer Activity in Professional Associations and Societies. *Voluntas*, 23(3):558–583, 2012.

M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Phys Rev E*, 68(3):36122, 2003.

K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *J Am Stat Assoc*, 96(455):1077–1087, 2001.

F. Ortega, J. M. Gonzalez-Barahona, and G. Robles. On the inequality of contributions to wikipedia. In R. H. Sprague Jr., editor, *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 304–310, Washington, DC, USA, 2008. IEEE Computer Society.

G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

G. Palla, A.-L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

P. Panzarasa, T. Opsahl, and K. M. Carley. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *J Am Soc Inf Sci Technol*, 60(5):911–932, 2009.

P. B. Reed and L. K. Selbee. The Civil Core in Disproportionality in Charitable Giving, Volunteering, Civic Participation. *Nonprofit Volunt Sect Q*, 30(4): 761–780, 2001.

S. G. B. Roberts, R. I. M. Dunbar, T. V. Pollet, and T. Kuppens. Exploring variation in active network size: Constraints and ego characteristics. *Soc Networks*, 31(2):138–146, 2009.

M. P. Rombach, M. A. Porter, J. H. Fowler, and P. J. Mucha. Core-Periphery Structure in Networks. *SIAM J Appl Math*, 74(1):167–190, 2014.

J. Saramäki, E. A. Leicht, E. López, S. G. B. Roberts, F. Reed-Tsochas, and R. I. M. Dunbar. Persistence of social signatures in human communication. *Proc Natl Acad Sci U S A*, 111(3):942–947, 2014.

V. Sekara, A. Stopczynski, and S. Lehmann. Fundamental structures of dynamic social networks. *Proc Natl Acad Sci U S A*, 113(36):9977–9982, 2016.

B. Wellman, J. Salaff, D. Dimitrova, L. Garton, M. Gulia, and C. Haythornthwaite. Computer networks as social networks: Collaborative work, telework, and virtual community. *Annu Rev Sociol*, 22(1):213–238, 1996.

J. Yang and J. Leskovec. Community-affiliation graph model for overlapping network community detection. In M. J. Zaki, A. Siebes, J. X. Yu, B. Goethals, G. Webb, and X. Wu, editors, *IEEE 12th International Conference on Data Mining (ICDM)*, pages 1170–1175, Los Alamitos, California, USA, 2012. IEEE Computer Society.

T. Zhang, P. Cui, C. Faloutsos, Y. Lu, H. Ye, W. Zhu, and S. Yang. comeNgo: A dynamic model for social group evolution. *ACM Trans Knowl Discov Data*, 11(4):1–22, 2017.