Followers Are Not Enough: A Multifaceted Approach to Community Detection in Online Social Networks

David Darmon[1,*]◑, Elisa Omodei[2,3]◑, Joshua Garland[4]◑,
1 Department of Mathematics, University of Maryland, College Park, Maryland, United States of America
2 LaTTiCe-CNRS, ISC-PIF, Paris, France
3 Department d'Enginyeria Informatica i Matematiques, Universitat Rovira i Virgili, Tarragona, Spain
4 Department of Computer Science, University of Colorado, Boulder, Colorado, United States of America
◑These authors contributed equally to this work.
*** E-mail: ddarmon@math.umd.edu (DD)**

Abstract

In online social media networks, individuals often have hundreds or even thousands of connections, which link these users not only to friends, associates, and colleagues, but also to news outlets, celebrities, and organizations. In these complex social networks, a 'community' as studied in the social network literature, can have very different meaning depending on the property of the network under study. Taking into account the multifaceted nature of these networks, we claim that community detection in online social networks should also be multifaceted in order to capture all of the different and valuable viewpoints of 'community.' In this paper we focus on three types of communities beyond follower-based structural communities: activity-based, topic-based, and interaction-based. We analyze a Twitter dataset using three different weightings of the structural network meant to highlight these three community types, and then infer the communities associated with these weightings. We show that interesting insights can be obtained about the complex community structure present in social networks by studying when and how these four community types give rise to similar as well as completely distinct community structure.

1 Introduction

Networks play a central role in online social media services like Twitter, Facebook, and Google+. These services allow a user to interact with others based on the online social network they curate through a process known as contact filtering [1]. For example, 'friends' on Facebook represent reciprocal links for sharing information, while 'followers' on Twitter allow a single user to broadcast information in a one-to-many fashion. Central to all these interactions is the fact that the *structure* of the social network influences how information can be broadcast or diffuse through the service.

Because of the importance of structural networks in online social media, a large amount of work in this area has focused on using structural networks for community detection. In this traditional view, a community is defined as a collection of nodes (users) within the network which are more highly connected to each other than to nodes (users) outside of the community [2, 3]. For instance, in [4], the authors use a follower network to determine communities within Twitter, and note that conversations tend to occur within these communities.

The approach of focusing on the structure of networks makes sense for 'real-world' sociological experiments, where obtaining additional information about user interactions may be expensive and time-consuming. However, with the prevalence of large, rich data sets from online social networks, additional information beyond the structure alone may be incorporated, and these augmented networks more realistically reflect how users interact with each other on social media services [5, 6].

A large body of work exists on methods for automatic detection of communities within networks, *e.g.,* the stochastic blockmodel [7] (and its recent generalization to weighted networks [8]), the Girvan-Newman algorithm [3], clique percolation [9], Infomap [10], Louvain [11], and the recently introduced OSLOM [12]. All these methods begin with a given network, and then attempt to uncover structure present in the network, *i.e.,* they are agnostic to how the network was constructed. As opposed to this agnostic analysis, we propose—and illustrate the importance of—a *multifaceted question-focused* approach. We believe that in order to understand all of the communities present in a network, it is important to take the following crucial steps:

1. Ask *several* questions about the communities that may be present in the network, each of which is aimed at revealing a new facet of the community structure present in the network—*e.g.,* which users are tightly connected structurally? which groups of users have similar topical interests?

2. Derive a weighting scheme aimed at answering each of the questions asked in the first step and then perform community detection on each weighted network.

3. Compare the (possibly) different communities that arise from this series of experiments on a micro and macro scale to unveil all the interesting facets of community structure present in the social network.

This is especially true for social network analysis. In social networks, a 'community' could refer to several possible structures. The simplest definition of community, as we have seen, might stem from the network of explicit connections between users on a service (friends, followers, etc.). On small time scales, these connections are more or less static. To form a more dynamic picture of community structure, we might instead determine communities based on who is talking to whom or which users talk about similar topics. More abstractly, we could even define communities as groups of people who exhibit similar activity profiles. We can characterize these types of communities based on the questions that motivate them:

- **Structure-based:** Who are your stated friends? Whom do you follow?

- **Activity-based:** Who shares similar activity profiles?

- **Topic-based:** What do you talk about?

- **Interaction-based:** Whom do you communicate with?

This is not meant to be an exhaustive list, but rather a list of some of the more common types of communities observed and studied in social networks. Instead of looking at each of these questions in isolation—which is the standard approach—we propose looking at when and how communities motivated by these different questions overlap and are disjoint, and whether different approaches to asking the question "What community are you in?" leads to different insights about a social network. For example, a user on Twitter might connect mostly with computational social scientists, utilize the service nearly every time a particular user or group of users is active, talk mostly about machine learning, and interact solely with close friends (who may or may not be computational social scientists). For this user, the answers to each question therefore correspond to more or less distinct communities. By contrast, a teenage user may only connect and interact with their friends, will most likely exhibit similar activity patterns dictated by the academic and extracurricular schedule of a student, and will discuss topics typical of their demographic. For this user, the answers to each question map to the same community of users, even though different definitions of community are being used. This highlights an important subtlety of this work: defining communities in different ways does not imply that the collection of users in each community *must* be different. Said differently, it is not obvious—or always the case—that changing the definition of community will always change the communities that are detected. Studying when this change does and does not occur—as we will show—provides interesting insights into the complex community structure present in these networks.

We now consider in more detail how each of these types of questions describes a unique facet of 'community'. Later in Section 4 we will show how asking *all* of these questions—instead of focusing on only one—gives a much deeper view of the (possibly) different communities that users belong to.

Naïvely the activity-based approach is motivated by the question of "Which users in a network have similar activity profiles?". With this question in mind, a community can then be thought of as those users who use (or do not use) the service at similar times. However, the measure we use is slightly more subtle that this. In particular, these communities can be thought of as groups of users who possess so-called "activity-based predictive capacity" with one another, *i.e.*, given a user u and follower f, who are members of a community detected with this weighting, there exists a reduction in uncertainty about the activity profile of follower f (tweeting or remaining silent), given the tweet history of user u, ignoring the information present in the tweet history of follower f. To accomplish this, we consider each user on Twitter as an information processing unit, but completely ignore the *content* of their tweets. We then weight the directed edges (the reported follower-followee relationships) between users with the so-called *transfer entropy* [13], calculated between the tweet history of follower f and user u. This is discussed in more detail in Section 2.3.

The topic-based and interaction-based approaches, in contrast to the activity-based approach, rely on the *content* of a user's interactions and ignore their temporal components. The content contains a great deal of information about communication between users. There are two broad approaches to topic-based communities in the literature. [14] used a set of users collected based on their use of a single hashtag, and tracked the formation of follower and friendship links within that set of users. In [15], the authors chose a set of topics to explore, and then seeded a network from a celebrity chosen to exemplify a particular topic. Both approaches thus begin with a particular topic in mind, and perform the data collection accordingly. Other approaches use probabilistic models for the topics and treat community membership as a latent variable [16]. For example, a popular approach to analyzing social media data is to use Latent Dirichlet Allocation (LDA) to infer topics based on the prevalence of words within a status [17, 18]. The LDA model can then be used to infer distributions over latent topics, and the similarity of two users with respect to topics may be defined in terms of the distance between their associated topic distributions. Because our focus is not on topic identification, we apply a simpler approach using hashtags as a proxy for topics, similar to the approaches presented in [19, 20]. We can then define the similarity of two users in terms of their hashtags, and use this similarity to build a topic-based network. This method is described in detail in Section 2.4.

Finally, the interaction-based approach relies on the meta-data and text of messages to identify whom a user converses with on the social media service. On Twitter, we can use mentions (indicating a directed communication) and retweets (indicating endorsement of another user) to identify conversation. A few works have investigated this type of community. For example, [21] considered both mention and retweet networks in isolation for a collection of users chosen for their political orientation. In [22], the authors construct a dynamic network based on simple time-windowed counts of mentions and retweets, and use the evolution of this network to aid in community detection.

Previous research on communities in social networks focused almost exclusively on different network types in isolation. A notable exception to this is [23], which considered both structure-based and interaction-based communities on Twitter. However, this study focused on data collected based on particular topics (country music, tennis, and basketball), and not on a generic subpopulation of Twitter users. Moreover, it did not explore the differences in community structure resulting from the different network weightings, and focused on aggregate statistics (community size, network statistics, etc.). Another notable exception is [24], where the authors used a tensor representation of user data to incorporate retweet and hashtag information into a study of the social media coverage of the Occupy Movement. The tensor can then be decomposed into factors in a generalization of the singular value decomposition of a matrix, and these factors can be used to determine 'salient' users. However, unlike our work, this approach focused on data for a particular topic (the Occupy Movement) and did not collect users based on a structural network.

Studying how the communities derived from follower-followee, activity-based, topic-based, and interaction-based networks are disjoint as well as overlapping allow for a more complete picture of the *implicit* networks present in online social media, as opposed to the *explicit* social network indicated by follower links alone. In this paper, we explore the relation between these various possible networks through their corresponding communities. We begin by describing the methodologies used to generate the various types of networks, and infer their community structure. We then explore how the communities of users differ depending on the type of network used. We conclude by illustrating that this multifaceted question-oriented approach to community detection provides interesting insights into the intricate multifaceted structure of several different users' communities, information that would not be available if only a single form of community detection was performed.

2 Methodology

In the following sections, we introduce the problem of community detection, and present the data set used for our analyses. We then describe our methodology for constructing the question-specific networks. In particular, we introduce an information theoretic statistic for activity-based communities, a retweet-mention statistic for interaction-based communities, and a hashtag similarity metric for defining topic-based communities. It should be noted that our goal is to choose representative weightings for each community type so that we may compare community structure across types, and not to—aside from the activity-based weighting—introduce novel weighting schemes *per se*.

2.1 Community Detection

As discussed in the introduction, we adopt the standard definition of *community*: a collection of nodes (users) within a network who are more densely connected to each other than with the rest of the network. Structural community detection is a well studied problem and several different methods and algorithms have been proposed. For a complete review of this subject we refer the reader to [25, 26]. In this paper however we focus on a class of networks and communities that is far less studied, in particular we study networks which are both *weighted* and *directed* and communities within those weighted directed networks that can (but need not) *overlap*. When selecting a detection algorithm we propose that all three (weight, direction, and overlap) are important for the following reasons. First, communication on Twitter occurs in a directed manner, with users broadcasting information to their followers. An undirected representation of the network would ignore this fact, and could lead to communities composed of users who do not actually share information. Second, we are interested in not just the structure of links but also in their function, and to capture this we use edge weightings which must be incorporated into the community detection process. Finally, since people can belong to multiple and possibly overlapping social (e.g., college friends, co-workers, family, etc.) and topical (e.g., a user can be interested in both cycling and politics and use the network to discuss the two topics with two different groups of users) communities, we are interested in finding *overlapping* communities, rather than partitions of the weighted directed network.

This last criterion in particular poses a problem because the majority of community detection algorithms developed so far are built to find partitions of a network and few are aimed at finding overlapping communities [8, 9, 27–33]. Among these methods, even fewer deal with directed or weighted networks. For example, the work of [9] on clique percolation can account for both features, but not at the same time. A recent method proposed in [12], OSLOM (Order Statistics Local Optimization Method), is one of the first methods able to deal with all of these features simultaneously. This method of Lancichinetti *et al.* relies on a fitness function that measures the statistical significance of clusters with respect to random fluctuations, and attempts to optimize this fitness function across all clusters. Following [12], the significance measure is defined as the probability of finding the cluster in a network without community structure. The null model used is highly similar to the one adopted by Newman and Girvan in [3] to define modularity, *i.e.,* it is a model that generates random graphs with a given degree distribution. The authors tested their algorithm on different benchmarks (LFR and Girvan-Newman) and real networks (such as the US air transportation network and a word association network), and compared its performance against the best algorithms currently available (*i.e.,* the ones mentioned in the introduction), and found excellent results. Moreover, they showed that OSLOM is also able to recognize the absence, and not simply the presence, of community structure, by testing it on random graphs. This feature of the algorithm plays an important role in the analysis of real world networks, since it is not always the case that community structure is indeed present and it is therefore useful to be able to detect its absence too. Therefore, given its versatility and performance with benchmark networks, in this paper we used OSLOM to detect *overlapping* communities present in our *weighted* and *directed* networks.

2.2 The Initial Dataset and Network Construction

The dataset for this study consisted of the tweets of 15,000 Twitter users over a 9 week period (from April 25th to June 25th 2011). The network and user activity (including tweets, mentions, and hashtags) were accessed using the Twitter API. The users are embedded in a network collected by performing a breadth-first expansion from a random seed user. In particular, once the seed user was chosen, the network was expanded to include his/her followers, but only included users considered to be 'active' (*i.e.,* users who tweeted at least once per day over the past one hundred days). Network collection continued in this fashion by considering the active followers of the active followers of the seed, and so on until 15,000 users were added to the network. The only biasing steps in this procedure are the selection of the seed node and the filtering out of inactive users. However, because all of the community types we are interested in involve some aspect of the interaction between users, each other, and their content, this procedure provides a reasonable follower-followee network for use with the different question-based analyses.

Since our goal is to explore the functional communities of this network, we filter the network down to the subset of users that actively interact with each other (*e.g.,* via retweets and mentions). We do this by measuring what we call (incoming/outgoing) information events. We define an outgoing information event for a given user u as either a mention made by u of another user in the network, or a retweet of one of u's tweets by another user in the network. The logic for this definition is as follows: if u mentions a user v this can be thought of as u directly sending information to v, and if u is retweeted by v then v received information from u and rebroadcast it to their followers. In either case there was information outgoing from u which affected the network in some way. Analogously, we define the incoming information event for u as either being mentioned by a different user in the network, or as retweeting another user in the network. With (incoming/outgoing) information events defined we filtered the network by eliminating all users with less than a total of 9 outgoing *and* incoming information events, *i.e.,* less than one information event per type per week on average. We then further restricted our analysis to the strong giant connected component of the network built from the (incoming/outgoing) information filtered set of users. In this study the link is directed from the user to the follower because this is the direction in which information flows. Thus, for a pair of users u and v, an edge $a_{v \to u}$ in the follower-followee network has weight 1 if user u follows v, and 0 otherwise. The final network consists of 6,917 nodes and 1,481,131 edges.

2.3 Activity-Based Communities and Transfer Entropy

For the activity-based communities, we consider only the timing of each user's tweets and ignore any additional content. From this starting point, we can view the behavior of a user u on Twitter as a point process, where at any instant t the user has either emitted a tweet ($X_t(u) = 1$) or remained silent ($X_t(u) = 0$). This is the view of a user's dynamics taken in [34] and [35]. Thus, we reduce all of the information generated by a user on Twitter to a time series $\{X_t(u)\}$ where t ranges over the time interval for which we have data (9 weeks in this case). Because status updates are only collected in discrete, 1-second time intervals, it is natural to consider only the discrete times $t = 1\text{s}, 2\text{s}, \ldots$, relative to a reference time.

Operationally, we expect users to interact with Twitter on a human time scale, and thus the natural one-second time resolution is too fine, since most humans do not write tweets on the time scale of seconds. For this reason, we coarsen each time series by considering non-overlapping time intervals ten minutes in length. For each time interval, we record a 1 if the user has tweeted during that time interval, and a 0 if he or she has not. Thus, the new coarsened time series now captures whether or not the user has been active on Twitter over any given ten minute time interval in our data set. It should be noted that, as (most) users are not constantly tweeting, these time series are quite sparse. For this particular application, this sparsity appropriately reflects the behavior of the users. However, this sparsity may not be appropriate for all applications; anytime the following measure is used in practice the effects of binning, sparsity, and their role in the activity-based community structure should be considered. We can then compute the flow of information from a user u to a follower f by computing the transfer entropy between their time series $X_t(u)$ and $X_t(f)$. See Appendix A for a detailed introduction to transfer entropy and its estimation from data.

For the communities based on transfer entropy, we weighted each edge from a user u to a follower f by the estimated transfer entropy of the user u on f,

$$w_{u \to f}^{\mathrm{TE}(k)} = \widetilde{\mathrm{TE}}_{X(u) \to X(f)}^{(k)}. \tag{1}$$

Transfer entropy is an information theoretic measure of *directed* information flow. We assume—as is standard—a positive transfer entropy from the tweet history of a user u to $X(f)$ implies a reduction in uncertainty in the activity of the follower, given the information contained in $X(u)$, while removing the information already contained in $X(f)$. This provides an information theoretic framework in which we can capture communities with activity-based predictive capacity between users or users with similar activity profiles. We call this relation "APC" (activity-based predictive capacity).

We computed the transfer entropy on each coarsened time series with lag k ranging from 1 to 6, this corresponds to a lag of ten minutes to an hour. The choice of lag must balance a trade-off between additional information and sparsity of samples: as we increase the lag k, we account for longer range dependencies, but we also decrease the number of samples available to infer a higher dimensional predictive distribution. Interestingly, as we will show later, the underlying communities resulting from the different lags have similar structure. See S1 Text for a discussion of these issues.

It should be noted that according to [35] a positive transfer entropy between $X(u)$ and $X(f)$ indicates that u "influences" f, or that u and f share a common influencer. Similarly, since transfer entropy is a nonlinear generalization of Granger causality [36], it is common to assume a positive transfer entropy between a user u and a follower f implies the relationship is causal in the Granger sense. However, we are skeptical whether this measure truly captures social influence in its entirety, or even causality in the case of online social networks. So instead we simply treat this weighting as a way to explicitly quantify APC between users.

As an aside, while we are skeptical that this measure can capture social influence, later in this paper we show that this information theoretic measure agrees with other concepts of influence such as the Forbes list of "Top 10 Social Media Influencers" and corroborates with the so-called strength of weak ties [37]. Even so, we do not explicitly treat transfer entropy—and we do not believe it should be treated without further investigation—as a quantification of social influence or causality. Instead this measure should be treated merely as a way to quantify APC between users.

To the best of our knowledge, this is the first use of transfer entropy for community detection in online social networks. Various other information theoretic approaches have been used successfully to analyze online social networks, *e.g.,* to gain insight into local user behavior [34], to detect communities based on *undirected* information flow [38], and to perform network inference and link prediction [35].

2.4 Topic-Based Communities and Hashtag Weighting

In contrast to the activity-based approach, the topic-based (or topical) communities, *i.e.,* communities defined by the topics users discuss, rely on the *content* of a user's interactions and ignore their temporal components. In a topical community, users are defined to be a member of a community if they tweet *about* similar topics as the other members of the community. In order to detect the topical communities, we weight the edges of the user-follower network through a measure based on the number of common hashtags between pairs of users. Hashtags are a good proxy for a tweet's content as hashtags are explicitly meant to be keywords indicating the topic of the tweet. Moreover they are widely used and straightforward to detect.

To this end, we characterize each user u by a vector $\vec{h}(u)$ of length equal to the number of unique hashtags in the dataset, and whose elements are defined as

$$h_i(u) = \phi_i(u) \log \frac{N}{n_i}, \tag{2}$$

where $\phi_i(u)$ is the frequency of hashtag i occuring in user u's tweets, N is the total number of users, and n_i is the number of users that have used the hashtag i in their tweets. This adapted term frequency–inverse document frequency (tf-idf) measure [39] captures the importance of a hashtag in the users's tweets through the first factor, but at the same time smooths it through the second factor by giving less importance to hashtags that are too widely used (as $\frac{N}{n_i}$ approaches one, its logarithm approaches zero).

For the topical communities we weight each directed edge from a user u to a follower f with the cosine similarity of their respective vectors $\vec{h}(u)$ and $\vec{h}(f)$:

$$w^{\mathrm{HT}}_{u \to f} = \frac{\vec{h}(u) \cdot \vec{h}(f)}{\left|\left|\vec{h}(u)\right|\right| \ \left|\left|\vec{h}(f)\right|\right|}. \tag{3}$$

This weight captures the similarity between users in terms of the topics discussed in their tweets.

2.5 Interaction-Based Communities and Mention / Retweet Weighting

Retweets and mentions are two useful features of Twitter networks which can be used to track information flow through the network. With mentions, users are sending direct information to other users and with retweets users are rebroadcasting information from a user they follow to all of their followers. This type of information flow defines a community in a much different way than transfer entropy. Instead of defining communities by the loss of uncertainty in one user's tweeting history based on another's, we define interaction-based communities by weighting the user-follower network with a measure proportional to the number of mentions and/or retweets between users. We define three weighting schemes. We first consider the number of tweets of a user u that were rebroadcast by a follower f, indicating a flow of information from u to f. Since different users produce different amounts of tweets and retweets, in order to take into account the relative importance of this measure we normalize it by dividing over the total number of retweets made by f,

$$w^{\mathrm{R}}_{u\to f} = \frac{\text{\# retweets of } u \text{ by } f}{\text{\# total retweets made by } f}, \tag{4}$$

We then consider the number of mentions of f made by user u. To account for the fact that some users/accounts are very popular and are often mentioned by so many other users that the flow of information from the latter to the former is flushed out, we normalize this measure by dividing over the total number of mentions received by f.

$$w^{\mathrm{M}}_{u\to f} = \frac{\text{\# mentions of } f \text{ by } u}{\text{\# total mentions of } f}, \tag{5}$$

Finally, we define the mention-retweet proportion as the arithmetic mean of the two measures just defined,

$$w^{\mathrm{MR}}_{u\to f} = \frac{(w^{\mathrm{R}}_{u\to f} + w^{\mathrm{M}}_{u\to f})}{2}. \tag{6}$$

3 Results and Discussion

In this section we show the importance and usefulness of a multifaceted approach to community detection in online social networks. We begin by showing that the communities emerging from the different weightings of the structural network quantitatively differ both at the macroscopic (*e.g.,* number of communities and their size distribution) and microscopic (*e.g.,* specific memberships, comemberships) scale in interesting ways. Finally, in order to provide a practical illustration of the utility of this question-oriented multifaceted approach, we present several concrete examples of multifaceted community memberships discovered with this method. We will show examples both at the community and individual user level.

3.1 Comparing Aggregate Statistics of Community Structure

We begin by examining the overall statistics for the communities inferred by OSLOM using the weightings defined in the previous sections. The number of communities by community type is given in Table 1. We see that the topic- and interaction-based networks admit the most communities. The activity-based network admits the fewest communities. One advantage of OSLOM over many other community detection algorithms is that it explicitly accounts for singleton 'communities': those nodes who do not belong to *any* extant communities. This is especially important when a network is collected via a breadth-first search, as in our network, where we begin from a seed node and then branch out. Such a search, once terminated, will result in a collection of nodes on the periphery of the network that may not belong to any community in the core.

We see in Table 1 that the topic- and interaction-based communities have the most singletons. This result for the interaction-based community is partially an artifact of the retweet/mention weighting: 717 of the users were disconnected from the network by how the weights were defined, resulting in 'orphan' nodes which we have included in the collection of singletons for all of our analyses. However, even after accounting for this artifact, the interaction-based network still has the most non-orphan singletons. This seems to indicate that a large fraction of the 6917 (nearly 25%) do not interact with each other in a concerted way that would mark them as a community under our interaction-based definition. This agrees with a result previously reported in [40] about how most users passively interact with incoming information on Twitter.

Table 1. Number of non-singleton communities and singletons by community type: S(tructural), A(ctivity-based), T(opic-based), and I(nteraction-based).

Community Type	# of Communities	# of Singletons
S	201	308
A, Lag 1	101	951
A, Lag 2	99	600
A, Lag 3	106	611
A, Lag 4	105	668
A, Lag 5	107	632
A, Lag 6	106	642
T	289	1064
I	252	2436

Next we consider the distribution of community sizes across the community types. The complementary cumulative distribution of community sizes is given in Fig. 1. Note that both axes are plotted on log-scales. Thus, for a fixed community size s, Fig. 1 shows the proportion of communities of size greater than s for each community type. We see that the community distributions have longer tails for the non-structural networks, and that the interaction-based network has the longest tail. Note that the activity-based communities, using transfer entropy estimated with varying lags, seem to converge on a similar large-scale community structure around lag 3. That is, the communities based on lag 1 and lag 2 transfer entropies are generally larger, and these communities resolve into smaller communities as the lag increases beyond 2. The distribution of community sizes for lag larger than 2 is insensitive to the lag. Most importantly, we see that the distributions of community sizes differ across the community types, the large-scale community structure is highly dependent on the particular question posed.

Figure 1. The proportion of communities greater than s in size, across the different community types. Note the logarithmic scale on the horizontal and vertical axes.

3.2 Comparing Community Structure with Normalized Mutual Information

In the previous section, we saw that the large scale statistics of the communities were highly dependent on the type of community under consideration. However, macroscale network statistics do not account for differences in community structure that result from operations such as splitting or merging of communities. Moreover, this view does not account for which users belong to which communities, and in particular which users belong to the same communities across community types. To answer this question, we invoke methods for the comparison of clusters: given two different clusterings of nodes into communities, how similar are the two clusters? The standard approach to answering this question is to define a metric on the space of possible partitions. Because we detect coverings rather than partitions, standard cluster comparison metrics like variation of information [41] are not appropriate. Instead, we use a generalization of variation of information first introduced in [31], the normalized mutual information. The normalized mutual information stems from treating clustering as a community identification problem: given that we know a node's community membership(s) in the first covering, how much information do we have about its community membership(s) in the second covering, and vice versa? Consider the two coverings $\mathcal{C}_1$ and $\mathcal{C}_2$. We think of the community memberships of a randomly chosen node in $\mathcal{C}_1$ as a binary random vector $\mathbf{X} \in \{0,1\}^{|\mathcal{C}_1|}$ where the i^{th} entry of the vector is 1 if the node belongs to community i and 0 otherwise. Similarly, $\mathbf{Y} \in \{0,1\}^{|\mathcal{C}_2|}$ is a binary random vector indicating the community memberships of the node in $\mathcal{C}_2$. Then the normalized mutual information is defined as

$$\mathrm{NMI}(\mathcal{C}_1, \mathcal{C}_2) = 1 - \frac{1}{2}\left(\frac{H[\mathbf{X}|\mathbf{Y}]}{H[\mathbf{X}]} + \frac{H[\mathbf{Y}|\mathbf{X}]}{H[\mathbf{Y}]}\right) \tag{7}$$

where $H[\cdot]$ denotes a marginal entropy and $H[\cdot|\cdot]$ denotes a conditional entropy. The normalized mutual information varies from 0 to 1, attaining the value of 1 only when $\mathcal{C}_1$ and $\mathcal{C}_2$ are identical coverings up to a permutation of their labels. See the appendix of [31] for more details.

We considered the normalized mutual information between the communities inferred from the structural network and the networks weighted with lag 1 through 6 transfer entropies, hashtag similarity, and mention, retweet, and mention-retweet activity. The resulting $\mathrm{NMI}(\mathcal{C}_i, \mathcal{C}_j)$ are shown in Fig. 2. Note the block diagonal structure in Fig. 2, which indicates that coverings are most similar within a question category. For example, the activity-based transfer entropy coverings are more similar to each other than to any of the other coverings. Similarly for the interaction-based mention, retweet, and mention-retweet coverings. This point may seem obvious, but the fact that the different weightings within a question category result in similar community structure indicates that each covering is capturing true, latent properties of the social network. Interestingly, the coverings resulting from the different weightings are all more similar to each other than to the structural covering from the unweighted network. Also note that the covering based on the hashtag similarities are different from all of the other weight-based coverings. In agreement with the results reported in the previous section, we see that the activity-based communities share similar structure for lags greater than 2. Because of these two results, in the remainder of the paper, we focus on the activity-based communities inferred using the lag-4 transfer entropy.

Figure 2. The normalized mutual information between the coverings inferred from the different community types. Values of normalized mutual information close to 1 indicate similarity in the community structure, while values close to 0 indicate dissimilarity. The normalized mutual information is computed with singletons and orphan nodes included. Note the block-diagonal structure, indicating the strong relationship between question type and community membership.

Thus, we see that although the activity-based, interaction-based, and topic-based communities relied on the structural network, their community structure differs *the most* from the community structure of the follower network. This agrees with the results from the previous section, and reinforces that the follower network is a necessary but not sufficient part of detecting communities characterized by properties beyond follower-followee relationships.

3.3 Comparing Edges Across Different Community Types

Any covering determined by OSLOM induces a natural partition of the edges in a directed network. In particular, let u and f be two users in the network, and let M_u and M_f be their community memberships. Then any edge $e_{u \to f}$ can be partitioned into one of three classes by

$$T(e_{u\to f}) = \begin{cases} \text{Inter-edge} & : \quad M_u \cap M_f = \emptyset \\ \text{Intra-edge} & : \quad M_u = M_f \\ \text{Mixed-edge} & : \quad \text{otherwise} \end{cases} . \tag{8}$$

In words, an inter-edge connects two users who share no community memberships, an intra-edge connects two users who each belong to the same communities, and a mixed-edge connects two users who share some, but not all, community memberships. Thus, inter-edges cross community boundaries, intra-edges lie within community boundaries, and mixed-edges lie at the borders of community boundaries. See Fig. 3 for a schematic of this edge partitioning.

Figure 3. A schematic of how, given a covering, the edges of the network can be partitioned using (8) into inter-edges, intra-edges, and mixed-edges. Inter-edges (red dashed) cross community boundaries. Intra-edges (blue solid) remain inside community boundaries. Mixed-edges (purple dotted) both remain in and cross community boundaries due to overlap in community membership. Each column corresponds to the same collection of weights, but partitioned using a different covering. Each row corresponds to the same covering, but for the different weights.

Each community type (Structural, Activity-based, Topic-based, Interaction-based) induces a different partition of the edges, and each edge type (Transfer Entropy, Hashtag, Mention-Retweet) induces a different distribution of weights. That is, let $W_{u \to f}$ be the weight on an edge $E_{u \to f}$ chosen uniformly at random from the network, and compute

$$P(W_{u \to f} \leq w_{u \to f} | T(E_{u \to f}) = t), \tag{9}$$

the empirical distribution over the edge weights conditioned on the edge type t being one of inter, intra, and mixed. The densities associated with these distributions are shown in Fig. 4. By considering how the weights vary conditional on the edge type, we can explore whether and how each of the follower-, interaction-, activity-, and topic-based communities define functional units within the network. For example, if users interact more-or-less independently of the topics they discuss, then the distributions of the interaction-based / topic-based edge weights should be independent of the edge type determined from the topic-based / interaction-based communities. By contrast, if users tend to interact mostly with those users who discuss similar topics, then the edge weight distributions should vary with the edge type.

Figure 4. The density of edge weights for different community types (rows) and weight types (columns). The red, blue, and purple values below each collection of densities indicate the median of weight on intra-, inter-, and mixed-edges, respectively.

Each *column* of Fig. 4 demonstrates how the density of weights change with the community type used to induce the partition. For example, the second column shows how the density of hashtag weights change based on the community partition. In general, the densities differ in non-trivial ways across the edge types and coverings. However, we see that the medians of the distributions provide a useful summary statistic. Under each collection of density functions, we list the median value of the inter, intra, and mixed-edges in red, blue, and purple, respectively. Unsurprisingly, we see that the greatest difference between the distributions occurs when we consider matching Community-type / Weight-type pairs, since the communities are determined based on the corresponding weights. However, we also see differences in the distributions when the Community-type / Weight-type pairs do not match. For example, for the covering corresponding to the activity-based communities, the hashtag weights tend to be higher for edges internal to communities than between communities. Similarly, for the covering corresponding to the topic-based communities, the transfer entropy weights tend to be higher on edges within (intra) and between (inter) communities, and lower for edges between members with multiple, non-identical memberships (mixed).

We also note that the distribution of transfer entropy weights always tend to be higher for edges crossing community boundaries (inter- or mixed-edges) compared to those edges within community boundaries (intra-edges). This is a property not exhibited by either the hashtag or mention-retweet weights, independent of the covering used to partition the edges. Moreover, for all but the activity-based covering, the weights of mixed-edges tend to be highest overall. Recall that the transfer entropy $\mathrm{TE}_{X(u) \to X(f)}$ quantifies the reduction in uncertainty about a follower

f's activity from knowing the activity of a user u. This result therefore implies that, in terms of prediction, it is more useful to know the time series of a user followed outside of the community compared to a user followed inside of the community, and even more useful to know the time series of a user that shares some, but not all, of the same community memberships. Thus, in an information theoretic sense, we see that novel information useful for prediction is more likely to flow *across* community boundaries than *within* community boundaries.

3.4 Uncovering the Multifacetedness of Community Memberships

In the previous sections we showed that the communities emerging from the different weightings of the structural network quantitatively differ both at the macroscopic and at the microscopic scale. Here, we present some concrete examples of different community membership, in order to provide a practical illustration of the utility of using a multifaceted approach to community detection. We explore this first on a community level and then at the individual level.

In the topic-based communities, we found a single community consisting of 83 users who tweet about environmental issues and frequently use hashtags such as #green, #eco and #sustainability. We also found a different community of 47 users who tweet about small businesses and entrepreneurship, using hashtags such as #smallbiz, #marketing and #enterpreneur. In both cases, almost all members of these topic-based communities are not found in the same community in the other networks *e.g.,* interaction- or activity-based, indicating that while these people talk about the same things and can therefore be considered a community based on their content, they do not strongly interact with each other nor behave the same, and so belong to different social groups with respect to interactions and behavior. This illustrates that the topics discussed by these users only define one facet of the users complex social network.

Another interesting example is a community whose topics tend to focus on Denver and Colorado. These users do not belong to the same community in the interaction-based network, but most of them do belong to the same community in the activity-based network. This indicates that these users react to the same events and issues regarding Colorado and are therefore strongly connected in the topic-based and activity-based networks, but at the same time they do not directly interact with each other and are therefore more loosely connected in the interaction-based networks, where they belong to different communities.

It is interesting to examine the intersection between this topic-based community and the overlapping activity-based communities and explore which users in the intersection of these community types have the highest APC. Users with the highest APC in an activity-based community would provide the highest reduction in uncertainty on average about that communities activity or inactivity on Twitter. Interestingly, among the highest APC users in these communities—those with highest transfer entropy on average—we find @Colorado, which is the state official Twitter account, @ConnectColorado, a page created to connect Coloradans, and the CBS Denver account, a popular news agency. This means that these accounts have high activity-based predictive capacity in terms of their followers' activity on Twitter. This is not surprising as these activity-based communities heavily overlapped with the topic-based communities that discuss Colorado (mentioned above) and these accounts provide information on this topic. This provides a brief proof of concept that the APC measure makes sense and suggests this could be a useful measure for identifying individuals who possess high activity-based predictive capacity for a communities activity on a social media network. However, more work needs to be done to quantify this more fully.

At the individual-based level, we also found interesting examples of multifaceted community membership, which show that asking multiple questions leads to finding different kinds of communities a user belongs to. For example, the user @JohnReaves belongs to two different topic communities: one about health care and the other about 'innovation' and 'creativity'. There was a 70% overlap between the innovation and creativity topic community, the interaction-based, and the activity-based community the user is in, whereas there is almost no overlap between the latter two and the healthcare community. This means that the user interacts with and behaves like the people talking about innovation and creativity, but even though he also talks about healthcare he does not interact with (nor behave like) the people talking about that. Another interesting example is given by user @jenajean. She belongs to two different topic communities: she talks about NASA and space but she does not interact nor behave like people belonging to that topic community, whereas she also talks about leadership, and the people she interacts with highly overlap with the people belonging to the topic-based community on leadership for which she belongs to. Lastly, @rajean, @Tekee, @Kimbirly and @mamamonroe are examples of users belonging to the topic community talking about Denver and Colorado. The activity-based community they belong to highly overlaps with the topic community (therefore they behave like the people that talk about Colorado), but they all belong to different interaction-based communities, each of which does not overlap in more than two users with the topic community.

4 Conclusion and Future Work

In this study, we have demonstrated that the communities observed in online social networks are highly question-dependent. The questions posed about a network *a priori* have a strong impact on the communities observed. Moreover, using different definitions of community reveal different and interesting relationships between users. More importantly, we have shown that these different views of the network are not revealed by using the structural network or any one weighting scheme alone. By varying the questions we asked about the network and then deriving weighting schema to answer each question, we found that community structure differed across community types on both the macro (*e.g.,* number of communities and their size distribution) and micro (*e.g.,* specific memberships, comemberships) scale in interesting ways.

To verify the validity of these communities we demonstrated that boundaries between communities represent meaningful internal/external divisions. In particular, conversations (*e.g.,* retweets and mentions) and topics (*e.g.,* hashtags) tended to be most highly concentrated within communities. We found this to be the case even when the communities were defined by a different criterion from the edge weights under study.

At first glance the boundaries defined by the activity-based communities derived from the transfer entropy weighting seemed less meaningful. However, upon further investigation our novel use of transfer entropy for the detection of activity-based communities highlighted an important fact about this social network: information transmission tended to be higher across community boundaries than within them. This result echos the ‘strength of weak ties’ theory from [37], which has found empirical support in online social networks [6]. This means that our use of transfer entropy not only defines boundaries that are meaningful divisions between communities but also illustrates that users who have a strong APC with a community need not be a member of that community.

Our findings may have important implications to a common problem in social network analysis: identification of influential individuals—but further work must be carried out. Many network measures of influence are based on the various types of centrality (degree, betweenness, closeness, eigenvector, etc.) [42]. Most centralities depend explicitly on the structure of the network under consideration. But we have seen in our study that the structural network only highlights one of many possible ways that users interact with online social media. Thus, a naïve application of centrality measures to a structural network may be answering a different question than the one motivating influence detection. For example, a user exhibiting high closeness centrality based on the structural network may be able to make a message visible to most other users in a social network, but this does not mean that those other users are likely to respond to or propagate that message. The importance of correctly framing network measures of influence has been explored previously [43], and our work further highlights its importance.

Based on our preliminary results, we conjecture that weighted generalizations of these centrality measures using transfer entropy might lead to better insights about who is actually influential in an online social network. It is interesting to note for example that two of the Forbes "Top 10 Social Media Influencers", happen to be in our network, *viz.*, Ann Tran and Jessica Northey, and transfer entropy also quantified these two users as having two of the highest APC on average, i.e., there was a high reduction in uncertainty about their followers activity on Twitter given their tweet histories. While we do not believe this information theoretic measure can capture social influence in its entirety, this suggests that this activity-based measure may be useful in finding influential members purely based on their temporal tweet history; even completely ignoring both tweet content and social status. However, a deeper analysis would need to be carried out to verify this finding. In addition to exploring this phenomenon further, we plan to explore a broader selection of choices for both the transfer-entropy lag and tweet history time resolution. We believe that by doing an in-depth analysis of both of these parameters we can discover interesting activity-based communities that occur on much broader time scales.

This work demonstrates the utility of a multifaceted question-oriented approach to community detection. This work shows that crafting *several* facet-driven weighting schema, doing community detection for each weighting scheme and then comparing the similarities and differences across community types, is an unavoidable process for uncovering the complex—and often hidden—community structure present in online social networks. More generally, this work illustrates that without a clear definition of community—or even only using a *single* definition of community—many rich and interesting communities present in online social networks remain invisible. Multifaceted question-oriented community detection can bring those hidden communities into the light.

Supporting Information

S1 Text

Transfer entropy and its estimation from data. Let $\{X_t\}$ and $\{Y_t\}$ be two strong-sense stationary stochastic processes[1]. In our work, these would correspond to the activities, $X_t(u)$ and $X_t(v)$, of two users u and v. We use the notation X_{t-k}^t to denote the values of the stochastic process from time $t-k$ to time t, $X_{t-k}^t = (X_{t-k}, X_{t-(k-1)}, \ldots, X_{t-1}, X_t)$. The lag-$k$ transfer entropy [13] of Y on X is defined as

$$\mathrm{TE}_{Y \to X}^{(k)} = H\left[X_t | X_{t-k}^{t-1}\right] - H\left[X_t | X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right], \tag{10}$$

where

$$H\left[X_t | X_{t-k}^{t-1}\right] = -E\left[\log_2 p(X_t | X_{t-k}^{t-1})\right] \tag{11}$$

and

$$H\left[X_t | X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right] = -E\left[\log_2 p(X_t | X_{t-k}^{t-1}, Y_{t-k}^{t-1})\right] \tag{12}$$

are the usual conditional entropies over the conditional (predictive) distributions $p(x_t | x_{t-k}^{t-1})$ and $p(x_t | x_{t-k}^{t-1}, y_{t-k}^{t-1})$. This formulation was originally developed in [13], where transfer entropy was proposed as an information theoretic measure of *directed* information flow. Formally, recalling that $H\left[X_t | X_{t-k}^{t-1}\right]$ is the uncertainty in X_t given its values at the previous k time points, and that $H\left[X_t | X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right]$ is the uncertainty in X_t given the joint process $\{(X_t, Y_t)\}$ at the previous k time points, transfer entropy measures the reduction in uncertainty of X_t by

[1] Recall that a stochastic process is strong-sense stationary if the joint distribution for the process evaluated at finitely many time points is invariant to an overall timeshift [44].

including information about Y_{t-k}^{t-1}, controlling for the information in X_{t-k}^{t-1}. By the 'conditioning reduces entropy' result [45]

$$H[X|Y,Z] \leq H[X|Y], \tag{13}$$

we can see that transfer entropy is always non-negative, and is zero precisely when

$$H\left[X_t|X_{t-k}^{t-1}\right] = H\left[X_t|X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right],$$

in which case knowing the past k lags of Y_t does not reduce the uncertainty in X_t. If the transfer entropy is positive, then $\{Y_t\}$ is considered causal for $\{X_t\}$ in the Granger sense [36].

When estimating transfer entropy from finite data, we will assume that the process $\{(X_t, Y_t)\}$ is jointly stationary, which gives us that

$$p(x_t|x_{t-k}^{t-1}) = p(x_{k+1}|x_1^k) \tag{14}$$

and

$$p(x_t|x_{t-k}^{t-1}, y_{t-k}^{t-1}) = p(x_{k+1}|x_1^k, y_1^k) \tag{15}$$

for all t. That is, the predictive distribution only depends on the past, not on when the past is observed. Given this assumption, we compute estimators for $p(x_{k+1}|x_1^k)$ and $p(x_{k+1}|x_1^k, y_1^k)$ by 'counting': for each possible marginal and joint past x_1^k and (x_1^k, y_1^k), we count the number of times a future of type x_{k+1} occurs, and normalize to obtain the appropriate estimators of the one-step-ahead predictive distributions. Call these estimators $\hat{p}(x_{k+1}|x_1^k)$ and $\hat{p}(x_{k+1}|x_1^k, y_1^k)$. Then the plug-in estimator for the transfer entropy is

$$\widehat{\text{TE}}_{Y\to X}^{(k)} = \hat{H}\left[X_t|X_{t-k}^{t-1}\right] - \hat{H}\left[X_t|X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right] \tag{16}$$

where we use the plug-in estimators $\hat{H}\left[X_t|X_{t-k}^{t-1}\right]$ and $\hat{H}\left[X_t|X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right]$ for the entropies. It is well known that the plug-in estimator for entropy is biased [46]. To account for this bias, we use the Miller-Madow adjustment to the plug-in estimator [47]. For a random variable X taking on finitely many values from an alphabet $\mathcal{X}$, the Miller-Madow estimator is

$$\tilde{H}[X] = \hat{H}[X] + \frac{|\hat{\mathcal{X}}| - 1}{2n} \tag{17}$$

where $|\hat{\mathcal{X}}|$ is the number of observed symbols from the alphabet $\mathcal{X}$ and n was the number of samples used to estimate $\hat{H}[X]$. The definition of transfer entropy (10) can be rewritten in terms of joint entropies as

$$\begin{aligned} TE_{Y\to X}^{(k)} &= H[X_t|X_{t-k}^{t-1}] - H[X_t|X_{t-k}^{t-1}, Y_{t-k}^{t-1}] && (18)\\ &= H[X_t, X_{t-k}^{t-1}] - H[X_{t-k}^{t-1}] - H[X_t, X_{t-k}^{t-1}, Y_{t-k}^{t-1}] + H[X_{t-k}^{t-1}, Y_{t-k}^{t-1}], && (19) \end{aligned}$$

We then apply the Miller-Madow adjustment individually to each of the entropy terms. For example, for the first term, we have

$$\tilde{H}[X_t, X_{t-k}^{t-1}] = \tilde{H}[X_{t-k}^t] = \hat{H}[X_{t-k}^t] + \frac{|\widehat{\mathcal{X}^{k+1}}| - 1}{2n}, \tag{20}$$

where $|\widehat{\mathcal{X}^{k+1}}|$ is the number of $(k+1)$-tuples we actually observe (of the 2^{k+1} possible tuples). Doing this for each term, the overall Miller-Madow estimator for the transfer entropy is

$$\begin{aligned} \widetilde{TE}_{Y\to X}^{(k)} &= \tilde{H}[X_t|X_{t-k}^{t-1}] - \tilde{H}[X_t|X_{t-k}^{t-1}, Y_{t-k}^{t-1}] && (21)\\ &= \tilde{H}[X_t, X_{t-k}^{t-1}] - \tilde{H}[X_{t-k}^{t-1}] - \tilde{H}[X_t, X_{t-k}^{t-1}, Y_{t-k}^{t-1}] + \tilde{H}[X_{t-k}^{t-1}, Y_{t-k}^{t-1}]. && (22) \end{aligned}$$

One possible problem with this estimator is that it can result in *negative* estimates of entropies. That usually occurs when $\hat{H}$ is very small. In these cases, we set the estimator to zero.

Acknowledgments

Thanks to Cesar Flores, Luís Seoane, Kevin Stadler, Jody Wright, and Nix Barnett for their contributions to preliminary ideas related to this paper and to Michelle Girvan, William Rand, Aaron Clauset, and Ryan James for their valuable comments and suggestions. Finally, we would like to acknowledge the Santa Fe Institute and its Complex Systems Summer School for providing the intellectually stimulating environment where this project began.

References

1. Cazabet R, Leguistin M, Amblard F (2012) Automated community detection on social networks: useful? efficient? asking the users. In: Proc. of the 4th WI&C. pp. 6:1–6:8.

2. Girvan M, Newman ME (2002) Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99: 7821–7826.

3. Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69: 026113.

4. Java A, Song X, Finin T, Tseng B (2009) Why we twitter: an analysis of a microblogging community. In: Advances in Web Mining and Web Usage Analysis, Springer. pp. 118–138.

5. Nguyen NP, Dinh TN, Xuan Y, Thai MT (2011) Adaptive algorithms for detecting community structure in dynamic social networks. In: Proc. of the IEEE 2011 INFOCOM. IEEE, pp. 2282–2290.

6. Grabowicz PA, Ramasco JJ, Moro E, Pujol JM, Eguiluz VM (2012) Social features of online networks: The strength of intermediary ties in online social media. PLoS ONE 7: e29358.

7. Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: First steps. Social networks 2: 109–137.

8. Aicher C, Jacobs AZ, Clauset A (2014) Learning latent block structure in weighted networks. To appear, J Compl Netw .

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

10. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci 105: 1118–1123.

11. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech 2008: P10008.

12. Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PLoS ONE 6: e18961.

13. Schreiber T (2000) Measuring information transfer. Phys Rev Lett 85: 461–464.

14. Rossi L, Magnani M (2012) Conversation practices and network structure in twitter. In: Proc. of the 6th ICWSM. pp. 563–566.

15. Lim KH, Datta A (2012) Following the follower: detecting communities with common interests on twitter. In: Proc. of the 23rd ACM HT. ACM, pp. 317–318.

16. Yin Z, Cao L, Gu Q, Han J (2012) Latent community topic analysis: integration of community discovery with topic modeling. ACM Trans Intell Syst Technol 3: 63:1–63:21.

17. Zhao WX, Jiang J, Weng J, He J, Lim EP, et al. (2011) Comparing twitter and traditional media using topic models. In: Proc. of the 33rd ECIR, Springer. pp. 338–349.

18. Michelson M, Macskassy SA (2010) Discovering users' topics of interest on twitter: a first look. In: Proc. of the 4th workshop on analytics for noisy unstructured text data. ACM, pp. 73–80.

19. Becker H, Naaman M, Gravano L (2011) Beyond trending topics: real-world event identification on twitter. In: Proc. of the 5th ICWSM. pp. 438–441.

20. Tsur O, Rappoport A (2012) What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In: Proc. of the 5th ACM WSDM. ACM, pp. 643–652.

21. Conover M, Ratkiewicz J, Francisco M, Gonçalves B, Menczer F, et al. (2011) Political polarization on twitter. In: Proc. of the 4th ICWSM. pp. 89–96.

22. Deitrick W, Hu W (2013) Mutually enhancing community detection and sentiment analysis on twitter networks. J Data Ana and Info Proc 1: 19–29.

23. Lim KH, Datta A (2012) Tweets beget propinquity: detecting highly interactive communities on twitter using tweeting links. In: Proc. of the IEEE/WIC/ACM WI-IAT. IEEE, pp. 214–221.

24. Kao A, Ferng W, Poteet S, Quach L, Tjoelker R (2013) Talison-tensor analysis of social media data. In: Intelligence and Security Informatics. IEEE, pp. 137–142.

25. Porter MA, Onnela JP, Mucha PJ (2009) Communities in networks. Notices of the AMS 56: 1082–1097.

26. Fortunato S (2010) Community detection in graphs. Phys Rep 486: 75–174.

27. Baumes J, Goldberg MK, Krishnamoorthy MS, Magdon-Ismail M, Preston N (2005) Finding communities by clustering a graph into overlapping subgraphs. In: Proc. of the IADIS AC. pp. 97-104.

28. Zhang S, Wang RS, Zhang XS (2007) Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Phys A 374: 483–490.

29. Gregory S (2007) An algorithm to find overlapping community structure in networks. In: Proc. of the 11th PKDD. pp. 91–102.

30. Nepusz T, Petróczi A, Négyessy L, Bazsó F (2008) Fuzzy communities and the concept of bridgeness in complex networks. Phys Rev E 77: 016107.

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

32. Evans TS, Lambiotte R (2009) Line graphs, link partitions, and overlapping communities. Phys Rev E 80: 016105.

33. Kovács IA, Palotai R, Szalay MS, Csermely P (2010) Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. PLoS ONE 5: e12528.

34. Darmon D, Sylvester J, Girvan M, Rand W (2013) Predictability of user behavior in social media: Bottom-up v. top-down modeling. In: ASE/IEEE Int'l Conf. on Social Computing. pp. 102–107.

35. Ver Steeg G, Galstyan A (2012) Information transfer in social media. In: Proc. of the 21st Inter. WWW Conf. ACM, pp. 509–518.

36. Granger CWJ (1963) Economic processes involving feedback. Inform Control 6: 28–48.

37. Granovetter MS (1973) The strength of weak ties. Am J Sociol 78: 1360–1380.

38. Darmon D, Omodei E, Flores CO, Seoane LF, Stadler K, et al. (2013) Detecting communities using information flow in social networks. In: Proc. of the CSSS, Santa Fe Institute.

39. Salton G, McGill MJ (1983) Introduction to modern information retrieval. New York: McGraw-Hill.

40. Romero DM, Galuba W, Asur S, Huberman BA (2011) Influence and passivity in social media. In: Machine learning and knowledge discovery in databases, Springer. pp. 18–33.

41. Meilă M (2003) Comparing clusterings by the variation of information. In: Learning theory and kernel machines, Springer. pp. 173–187.

42. Newman M (2009) Networks: an introduction. Oxford University Press.

43. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, et al. (2010) Identification of influential spreaders in complex networks. Nature Phys 6: 888–893.

44. Grimmett G, Stirzaker D (2001) Probability and random processes. Oxford University Press.

45. Cover TM, Thomas JA (2006) Elements of information theory. Wiley-Interscience.

46. Paninski L (2003) Estimation of entropy and mutual information. Neural Comput 15: 1191–1253.

47. Miller GA (1955) Note on the bias of information estimates. In: Quastler H, editor, Information theory in psychology II-B, Glencoe, Illinois: Free Press. pp. 95–100.