# Followers Are Not Enough: Beyond Structural Communities in Online Social Networks

David Darmon[1], Elisa Omodei[2], and Joshua Garland[3]

[1]University of Maryland, Dept. of Mathematics
[2] LaTTiCe (CNRS, ENS, Paris 3), ISC-PIF
[3] University of Colorado at Boulder, Dept. of Computer Science

April 22, 2014

**Abstract**

Community detection in online social networks is typically based on the analysis of the explicit connections between users, such as "friends" on Facebook and "followers" on Twitter. But online users often have hundreds or even thousands of such connections, and many of these connections do not correspond to real friendships or more generally to accounts that users interact with. We claim that community detection in online social networks should be question-oriented and rely on additional information beyond the simple structure of the network. The concept of 'community' is very general, and different questions such as "who do we interact with?" and "with whom do we share similar interests?" can lead to the discovery of different social groups. In this paper we focus on three types of communities beyond 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 the communities obtained in the three weighted cases are highly different from each other, and from the communities obtained by considering only the unweighted structural network. Our results confirm that asking a precise question is an unavoidable first step in community detection in online social networks, and that different questions can lead to different insights into the network under study.

# 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 [**?**]. 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 of 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. Here, by 'community' we mean the standard definition from the literature on social networks: a collection of nodes (users) within the network who are more highly connected to each other than to nodes (users) outside of the community [**?**]. For instance, in [**?**], 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 structural 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 for online social networks, additional information beyond the structure alone may be incorporated, and these augmented networks more realistically reflects how users interact with each other on social media services [**?**].

A large body of work exists on methods for automatic detection of communities within networks, among which we recall Girvann and Newman algorithm [32, 33], Infomap [39], Leuven [3], and the recently introduced OSLOM [24]. 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 question-focused approach. We believe that in order to understand the communities present in a data set, it is important to begin with a clear picture of the community type under consideration, and then perform the network collection and community detection with that community type in mind.

This is especially true for social network analysis. In online 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, and we might instead determine communities based on who is talking to whom, providing a more dynamic picture. On a more abstract level, a user might consider themselves part of a community of people discussing similar topics. We might also define communities as collections of people who exhibit similar behaviors on a service, as in communities of teenagers vs. elderly users. We can characterize these types of communities based on the types of questions we might ask about them:

- **Structure-based:** Who have you stated you are friends with? Who do you follow?
- **Activity-based:** Who do you act like?

- **Topic-based:** What do you talk about?
- **Interaction-based:** Who 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 in social networks. We propose looking at when and how communities motivated by these different questions overlap, 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, talk mostly about machine learning, interact solely with close friends (who may or may not be computational social scientists), and utilize the service only on nights and weekends. Each of these different 'profiles' of the user highlight different views of the user's social network, and represent different types of communities. We divide our approaches into four categories based on the questions outlined above: structure-based, activity-based, topic-based, and interaction-based. The structure-based approach, as outlined above, is most common, and for our data relies on reported follower relationships.

The activity-based approach is motivated by the question of which individuals act in a homogeneous manner, e.g., which users use a service in a similar way. The main tools for answering this question stem from information theory. We consider each user on an online social network as an information processing unit, but ignore the content of their messages. In particular, our current activity-based approach was originally motivated by a methodology used to detect functional communities within populations of neurons [**?**]. Similar information theoretic approaches have been used with data arising from online social networks to gain insight into local user behavior [**?**], to detect communities based on undirected information flow [**?**], and to perform link detection [**?**].

Our 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 the communication between users. 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 [**?**, **?**]. 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 [**?**, **?**]. We can then define the similarity of two users in terms of their hashtags, and use this similarity to build a topic-based network.

Finally, the interaction-based approach relies on the meta-data and text of messages to identify who 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. Moreover, we can define a directed influence between two users by considering the attention paid to that user compared to all other users. This allows us to generate a network based on conversations and user interactions.

The activity-based, topic-based, and interaction-based networks allow us to build a more complete picture of the *latent* social network present in online social media, as opposed to the *explicit* social network indicated by structural links. In this paper, we

explore the relation between these various possible networks and 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. Finally, we explore how communication patterns differ across and within the different community types.

## Related Work

Previous research on communities in social networks focused almost exclusively on different network types in isolation. For example, an early paper considered the communities, and associated statistics, inferred from a follower network on Twitter [**?**]. More recent work has considered the dynamics of communities based on structural links in Facebook [**?**] and how structural communities impact mentions and retweets on Twitter [**?**].

Information theoretic, activity-based approaches have been applied previously to the analysis of networks arising in online social media [**?**, **?**], but to the best of our knowledge this is the first use of transfer entropy, an information theoretic measure of directed influence, for community detection.

For interaction-based communities, [**?**] considered both mention and retweet networks in isolation for a collection of users chosen for their political orientation. In [**?**], 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.

There are two broad approaches to topic-based communities in the literature. [**?**] 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 [**?**], 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 [**?**].

A notable exception to the analysis of isolated types of communities is [**?**], 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 focuses on aggregate statistics (community size, network statistics, etc.). Another notable exception is [**?**], 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, this approach focused on data for a particular topic (the Occupy Movement) and did not collect users based on a structural network.

# 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 method for activity-based communities, a retweet-mention statistic for interaction-based communities, and a hashtag similarity metric for defining topic-based communities.

## 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 [**?**]. 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 the two different communities) 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 [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. Among these methods, even fewer deal with directed or weighted networks. For example, the work of [**?**] on clique percolation can account for both features, but not at the same time. A recent method proposed by [**?**], OSLOM (Order Statistics Local Optimization Method), is one of the first methods able to deal with all of these features simultaneously. Their method 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 [**?**], the significance measure is defined as the probability of finding the cluster in a network without community structure. The null model used is basically the same as the one adopted by Newman and Girvan 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 and the 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 data networks, since it is not always the case that a 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 network, in this paper we used OSLOM to detect *overlapping* communities present in our *weighted* and *directed* network.

## 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 users are embedded in a network collected by performing an intelligent 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.

Since our goal is to explore the functional communities of this network, we filter the network down to the subset of users which 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 the 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 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 the information (in the case of transfer entropy) or influence (in the case of mention-retweets) flows. Thus, for a pair of users $u$ and $v$, an edge $a_{v\to u}$ in the structural 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.

## 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 [**?**] and [**?**]. 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. We can then compute the flow of information between two users $u$ and $v$ by computing the transfer entropy between their time series $X_t(u)$ and $X_t(v)$.

Let $\{X_t\}$ and $\{Y_t\}$ be two strong-sense stationary stochastic processes. 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 of $Y$ on $X$ is defined as

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

where

$$H\left[X_t|X_{t-k}^{t-1}\right] = -E\left[\log_2 p(X_t|X_{t-k}^{t-1})\right] \tag{2}$$

and

$$H\left[X_t|X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right] = -E\left[\log_2 p(X_t|X_{t-k}^{t-1}, Y_{t-k}^{t-1})\right] \tag{3}$$

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

$$H[X|Y, Z] \leq H[X|Y], \tag{4}$$

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 [**?**].

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

$$p(x_t|x_{t-k}^{t-1}) = p(x_{k+1}|x_1^k) \tag{5}$$

and

$$p(x_t|x_{t-k}^{t-1}, y_{t-k}^{t-1}) = p(x_{k+1}|x_1^k, y_1^k) \tag{6}$$

for all $t$. That is, the predictive distribution only depends on the past, not on when the past is observed[1]. 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 past $(x_1^k, y_1^k)$, we count the number of times a future of type $x_{k+1}$ occurs, and normalize. 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

$$\widehat{\mathrm{TE}}_{Y\to X}^{(k)} = \hat{H}\left[X_t|X_{t-k}^{t-1}\right] - \hat{H}\left[X_t|X_{t-k}^{t-1}, Y_{t-k}^{t-1}\right] \tag{7}$$

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 [**?**]. To account for this bias, we use the Miller-Madow adjustment to the plug-in estimator [**?**].

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

$$w_{u\to f}^{\mathrm{TE}(k)} = \widehat{\mathrm{TE}}_{X(u)\to X(f)}^{(k)}. \tag{8}$$

A positive value for the transfer entropy of the user $u$ on $f$ indicates that $u$ influences $f$, or that $u$ and $f$ share a common influencer [**?**].

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. 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 they have 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. We then compute the transfer entropy on these coarsened time series taking $k$ to range from 1 to 6, which 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, we account for longer range dependencies, but we also decrease the number of samples available to infer a higher dimensional predictive distribution. Ultimately, due to similarities in the underlying communities we chose a lag-4 transfer entropy. All references to activity-based weights, unless otherwise noted, refer to this choice of lag.

## 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 directed 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. For

[1] We really only need *conditional* stationarity [**?**], but stationarity implies conditional stationarity.

the interaction-based communities we consider three weighting schemes: proportional retweets,

$$w^{\mathrm{R}}_{u\to f} = pR = \frac{\text{\# retweets of } u \text{ by } f}{\text{\# total retweets made by } f}, \tag{9}$$

proportional mentions,

$$w^{\mathrm{M}}_{u\to f} = pM = \frac{\text{\# mentions of } f \text{ by } u}{\text{\# total mentions of } f}, \tag{10}$$

and mention-retweet as the arithmetic mean of these two measures,

$$w^{\mathrm{MR}}_{u\to f} = \frac{(pM + pR)}{2}. \tag{11}$$

## Topic-Based Communities and Hashtag Weighting

The final community we consider is a topic-based or topical community, i.e., a community defined by the content (topics) users discuss. So in a topical community, users are defined to be a member of a community if they tweet *about* similar topics as 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

$$h_i(u) = n_i(u) * \log \frac{N}{n_i} \tag{12}$$

where $n_i(u)$ 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 [**?**] 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 $\vec{h}(f)$:

$$w^{\mathrm{HT}}_{u\to f} = \frac{\vec{h}(u)\cdot\vec{h}(f)}{||\vec{h}(u)||\ ||\vec{h}(f)||}. \tag{13}$$

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

# Results and Discussion

## 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 least number of communities. One advantage of the 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, with the activity-based community dominating this measure. This result for the activity-based community is partially an artifact of a property 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 [**?**] 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 (1719) |

Next we consider the distribution of community sizes across the community types. The complementary cumulative distribution of community sizes is given in Figure 1. Note that both axes are plotted on log-scales. Thus, for a fixed community size $c$, Figure 1 shows the proportion of communities of size greater than $c$ 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. The largest communities for the structural, activity-based, topic-based, and interaction-based networks have 198, 358, 338, and 811 members, respectively. Most importantly, we see that the distributions of community sizes differ across the community types, highlighting that the different networks give rise to different large-scale community structure dependent on the particular weighting of the structural network.

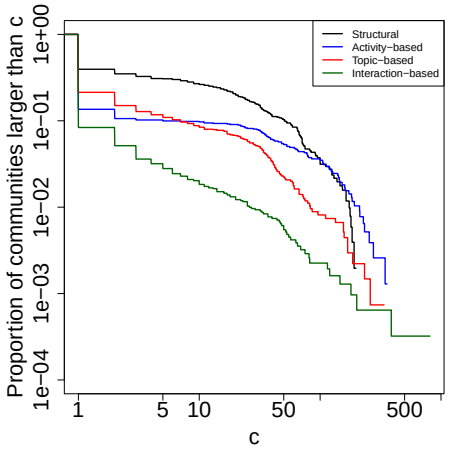


Figure 1: The proportion of communities greater than $c$ in size, across the different community types. Note the logarithmic scale on the horizontal and vertical axes.

## 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 [**?**] are not appropriate. Instead, we use a generalization of variation of information first introduced in [**?**], 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 $i^{\text{th}}$ 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

$$\text{NMI}(\mathcal{C}_1, \mathcal{C}_2) = 1 - \frac{1}{2}\left(\frac{H[\mathbf{X}|\mathbf{Y}]}{H[\mathbf{X}]} + \frac{H[\mathbf{Y}|\mathbf{X}]}{H[\mathbf{Y}]}\right) \tag{14}$$

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 [**?**] 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 $\text{NMI}(\mathcal{C}_i, \mathcal{C}_j)$ are shown in Figure 2. We see that similarity between the coverings is dictated by the generic community type (structural, activity-based, etc.). That is, the transfer entropy coverings are more similar to each other than to any of the other coverings, with a similar result for the mention, retweet, and mention-retweet coverings. 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.

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.

## Comparing Edges Across Different Community Types

We next explore how the edge weights defined by equations (8), (11), and (13), and thus different forms of information flow, differ between community types. For a fixed community type, edges for a particular community may be partitioned into three sets: those from a user in the community to another user in the community (internal-to-internal), those from a user in the community to a user outside of the community (internal-to-external), and those from a user outside the community to a user inside the community (external-to-internal). See Figure 3 for a schematic of this edge partitioning. For a meaningful community, we expect the distribution of weights within the community (internal-to-internal weights) to be different from the distribution of weights without the community (internal-to-external and external-to-internal).

As an example, Figure 4 shows the distributions of hashtag-based weights for the largest community in the mention-retweet network. We see that the distribution of internal-to-internal hashtag weights has a longer tail than either the external-to-internal or internal-to-external hashtag weights, with edges within the community having higher

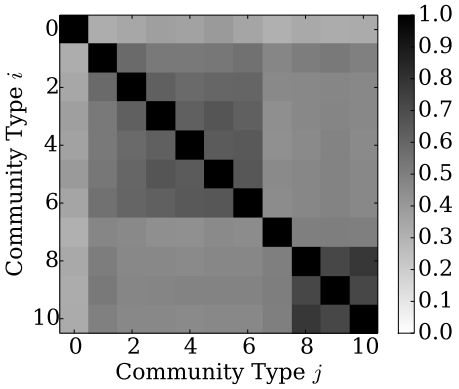


Figure 2: The normalized mutual information between the coverings inferred from the different community types. Community type 0 corresponds to the structural communities, community types 1 through 6 correspond to the activity-based communities with lag 1 through 6 transfer entropies, community type 7 corresponds to the topic-based communities, and community types 8, 9, and 10 correspond to the interaction-based communities using mentions, retweets, and both mentions and retweets. 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.

weights than edges crossing the boundary of the community. Thus, while the community was defined in terms of interactions, we still see a shift in the distribution of topic-similarity.

For all four community types, the transfer entropy tended to be higher for edges crossing community boundaries than for those internal to community boundaries. 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. 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.

Note that the communities defined by the follower network do tend to have higher edge weights internal compared to across community boundaries. Thus, we do see that the structural communities capture some information about the functional behavior of communities of users in terms of topics and interaction. However, the ratio is not as large as when we explicitly seek out communities based on a particular type of functional community. This again emphasizes the importance of properly formulating the goal of a community detection study in the context of online social networks.

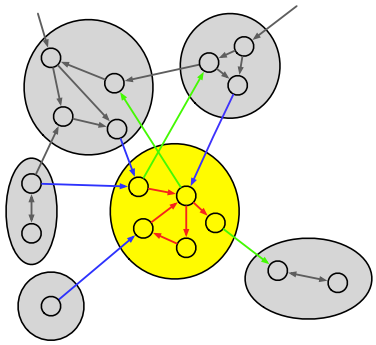

Figure 3: An example of the edges considered in determining the edge weight distribution for a given community (the focal community is in yellow). We focus on the internal-to-internal (red), internal-to-external (green), and external-to-internal (blue) edges. For a given focal community, all other edges (grey) are not considered.

## Qualitative Analysis of Community Memberships Across Types

As demonstrated by [**?**] in the context of modularity maximization-based community detection, an exponential number of nearby partitions may exist that nearly maximize an objective function used to measure the goodness-of-fit of a graph partition used for community detection. Because of this and related issues, it is always wise to perform some sort of qualitative study of the communities returned by any community detection algorithm to verify their meaningfulness with respect to the scientific question at hand. In this section, we consider a collection of communities in such a study.

In the topic-based communities, we find a single community consisting of 83 users who tweet about environmental issues and frequently use hashtags such as #green, #eco and #sustainability. We also find a different community of 47 users who tweet about small businesses and entrepreneurship, using hashtags such as #smallbiz, #marketing and #enterpreneur. In both cases most members of the topic-based communities are not found in the same community in the other networks, 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.

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. As expected, among the most influential users (in terms of transfer entropy) we find Colorado, which is the state official Twitter account, Con-

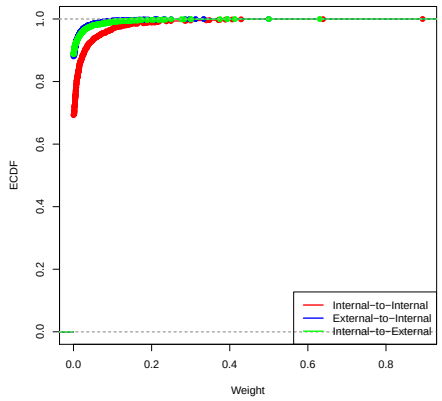


Figure 4: The proportion of edges with a weight at least as large as the weight on the horizontal axis, across the types of edges described in Figure 3. The community is defined by user interactions, and the edge weights are determined by topic similarity. The dashed vertical lines indicate the median weight for each type of edge. Note the logarithmic scale on the horizontal axis.

nectColorado, a page created to connect Coloradans, and CBS Denver account.

Lastly, it is interesting to note that in the top ten most influential users (ranked using the total outgoing strength in the activity-based network) we find two users (Ann Tran and Jessica Northey) that were listed by Forbes in the "Top 10 Social Media Influencers".

## 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: influence tended to be higher across community boundaries than within them. This result echos the 'strength of weak ties' theory from [**?**], which has found empirical support in [**?**] for online social networks. 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 influence on a community need not be a member of that community.

Our findings have important implications to a common problem in social network analysis: identification of influential individuals. Many network measures of influence are based on the various types of centrality (degree, betweenness, closeness, eigenvector, etc.) [**?**]. Most centralities depend explicitly on the structure of the network under consideration. But we have seen in our study that a structural network alone is not sufficient to capture user interaction or influence in online social media. Thus, a naïve application of centrality measures to a structural network for influence detection may give rise to erroneous results. This result has been explored previously [**?**], and our work further highlights its importance. We believe that weighted generalizations of these centralities using transfer entropy might lead to better insights about who is actually influential in an online social network. 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 that asking the proper question and then crafting an appropriate weighting scheme to answer that question is an unavoidable first step for community detection in online social media. More generally, this work illustrates that without a clear definition of community, many rich and interesting communities present in online social networks remain invisible. Question-oriented community detection can bring those hidden communities into the light.