

# 1 Introduction

Networks play a central role in online social networks like Twitter, Facebook, and Google+. These services allow a user to interact with other based on the online social network they curate, also 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 dictates the ways that information can be broadcast or diffuse through the service.

Because of the importance of structural networks, a large amount of work has focused on using structural networks to detect communities in online social media. Here, we use the term community to mean the standard definition of collection of nodes (users) within the network who are more highly connected to each than to users outside of the community [?]. For instance, in [?], the authors use a friendship 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 users 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 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 [?, ?, ?, ?, ?]. See [?] for a recent review. All these methods *begin* with a given network, and then attempt to uncover structure present in the network. That is, they are agnostic to how the network was constructed. Because of this, we propose that in answering any sort of question about the communities present in a data set, it is important to begin with a clear picture of the type of community under consideration, and then to tailor the construction of the network and the use of detection algorithms towards that goal.

This is especially true for social network analysis. In online social networks, ‘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. On a more abstract level, a user might consider themselves part of a community of people discuss similar topics. We might also define communities as collections of people who exhibit similar behaviors on a service, as in a 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 are you 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 talk to? Who do you talk about?

We propose looking at when and how communities motivated by 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 would be highlighted by a different view of the user’s social network. 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 concerted manner. 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. This viewpoint has been successfully applied to gain insight into local behavior in online social networks [?]. The information processing framework applies equally well to spatially extended systems. In particular, our current activity-based approach was originally motivated by a methodology used to detect functional communities within populations of neurons [?]. This approach has been extended to social systems, detecting communities on Twitter based on undirected information flow [?]. Others have successfully applied a similar viewpoint using transfer entropy, a measure of directed information flow, to perform link detection on Twitter [?]. However, the use of transfer entropy has not been used for community detection, and we extend this previous work to incorporate transfer entropy-based communities.

Our topic-based and interaction-based approaches, in contrast to the activity-based approach, rely on the *content* of a user’s interaction and ignore its temporal component. 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. The interaction-based approach relies on meta-data about 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 interaction.

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. **TK: Put a blurb here about the qualitative study.** Finally, we explore how the different communities relate in terms

## 2 Related Work

Most previous work on communities in online social networks have focused on these types of communities in isolation. For instance (**References from links.txt go here.**)

A notable exception to the analysis of different types of networks is [?], which considered both structure-based and interaction-based communities on Twitter. However

...

**TK: Look into the use of ‘multilevel’ / hybrid networks in social media.**