Working Title

immediate

December 11, 2013

Abstract

Lorem ipsum

1 Introduction

Networks play a central role in online social networks like Twitter, Facebook, and Google+. These services have become popular largely because of the fact that they allow for the interaction of millions of users in ways restricted by the network each user chooses to be part of. Central to this interaction is the concept

A large body of work exists on methods for automatic detection of communities within networks **(lots of references go here, starting with Newman and ending with Fortunato)**. 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, 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

This is especially true for social network analysis. In online social networks, 'community' could mean many things. The simplest definition of community 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:

- **Structural:** Who are you friends with?
- **Interaction-based:** Who do you talk to?
- **Topic-based:** What do you talk about?

- **Behavioral:** Who do you act like?

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.)**

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, **(Put a 'worked example' here as to how a single person might belong to various different types of communities, and how these types of communities might reveal different insights.)**

In particular, we can break down our approaches into two broad categories: content-free and content-full. The content-free 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 **(Cite Shannon here, with his ideas that the content of the message doesn't matter for information theory?)**. We have successfully used this viewpoint to gain insight into local behavior in online social networks [5]. The information processing framework applies to equally well to spatially extended systems. In particular, our current content-free approach was originally motivated by a methodology used to detect functional communities within populations of neurons [10]. We have extended this work to social systems, detecting communities on Twitter based on undirected information flow [4]. Others have successfully applied a similar viewpoint. For instance, in [11], the authors use transfer entropy, a measure of directed information flow, to perform link detection on Twitter.

In contrast to the content-free approach, a content-full approach would take into account the *content* being transmitted via an online social network. This content has a great deal of social information embedded in it. For example, on Twitter, a tweet might have a hashtag (indicating a topic), a mention or reply (indicating a directed communication), or a retweet (indicating endorsement of another user). This information allows us to build a more complete picture of the *latent* social network, as opposed to the *explicit* social network indicated by friend and follower links.

Many approaches have explicitly accounted for this information in their definition of a community. **(Put references to work done using mentions / replies / retweets / hash tags, etc.)**

Wrap up. Probably write this *after* more of our results are in.

2 Activity-Based v. Interaction-Based Communities

Lorem ipsum.

3 Methodology

3.1 Activity-Based Communities and Transfer Entropy

Suppose we have two stochastic processes $\{X_t\}$ and $\{Y_t\}$. Lag-k transfer entropy is defined as

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

where

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

and

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

are the usual conditional entropies over the conditional (predictive) distributions $p(x_t|x^{t-1}_{t-k})$ and $p(x_t|x^{t-1}_{t-k}, y^{t-1}_{t-k})$. This formulation was originally developed in [9], where transfer entropy was proposed as an information theoretic measure of *directed* information flow. Formally, recalling that $H\left[X_t|X^{t-1}_{t-k}\right]$ is the uncertainty in X_t given its values at the previous k time points, and that $H\left[X_t|X^{t-1}_{t-k}, Y^{t-1}_{t-k}\right]$ 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 by including information about Y_t. By the 'conditioning reduces entropy' result [3]

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

we can see that transfer entropy is always non-negative, and is zero precisely when $H\left[X_t|X^{t-1}_{t-k}\right] = H\left[X_t|X^{t-1}_{t-k}, Y^{t-1}_{t-k}\right]$, 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 [6, 1].

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

$$p(x_t|x^{t-1}_{t-k}) = p(x_{k+1}|x^k_1) \tag{5}$$

and

$$p(x_t|x^{t-1}_{t-k}, y^{t-1}_{t-k}) = p(x_{k+1}|x^k_1, y^k_1) \tag{6}$$

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 $p(x_{k+1}|x^k_1)$ and $p(x_{k+1}|x^k_1, y^k_1)$ by 'counting': for each possible past (x^k_1, y^k_1), we count the number of times a future of type x_{k+1} occurs, and normalize. Call these estimators $\hat{p}(x_{k+1}|x^k_1)$ and $p(x_{k+1}|x^k_1, y^k_1)$. Then the plug-in estimator for the transfer entropy is

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

where we use the plug-in estimators $\hat{H}\left[X_t|X^{t-1}_{t-k}\right]$ and $\hat{H}\left[X_t|X^{t-1}_{t-k}, Y^{t-1}_{t-k}\right]$ for the entropies. It is well known that the plug-in estimator for entropy is biased [8]. To account for this bias, we use the Miller-Madow adjustment to the plug-in estimator [7].

[1] We really only need *conditional* stationarity [2], but stationarity implies conditional stationarity

3.2 Interaction-Based Communities and Mention / Retweet Weighting

References

[1] Lionel Barnett and Terry Bossomaier. Transfer entropy as a log-likelihood ratio. *Physical Review Letters*, 109(13):138105, 2012.

[2] S Caires and JA Ferreira. On the nonparametric prediction of conditionally stationary sequences. *Probability, Networks and Algorithms*, (4):1–32, 2003.

[3] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[4] David Darmon, Elisa Omodei, Cesar O Flores, Luis F Seoane, Kevin Stadler, Jody Wright, Joshua Garland, and Nix Barnett. Detecting communities using information flow in social networks. In *Proceedings of the Complex Systems Summer School*. Santa Fe Institute, 2013.

[5] David Darmon, Jared Sylvester, Michelle Girvan, and William Rand. Understanding the predictive power of computational mechanics and echo state networks in social media. *HUMAN*, 2(1):pp–13, 2013.

[6] Clive William John Granger. Economic processes involving feedback. *Information and Control*, 6(1):28–48, 1963.

[7] George A Miller. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, 2:95–100, 1955.

[8] Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

[9] Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

[10] Cosma Rohilla Shalizi, Marcelo F Camperi, and Kristina Lisa Klinkner. Discovering functional communities in dynamical networks. In *Statistical network analysis: Models, issues, and new directions*, pages 140–157. Springer, 2007.

[11] Greg Ver Steeg and Aram Galstyan. Information transfer in social media. In *Proc. 21st Int'l World Wide Web Conf.*, pages 509–518. ACM, 2012.