QUALITY ENGINEERING
https://doi.org/10.1080/08982112.2018.1501063

Discussion of "Real-time monitoring of events applied to syndromic surveillance"

James D. Wilson

University of San Francisco, San Francisco, California

ABSTRACT
I discuss the article "Real-time monitoring of events applied to syndromic surveillance" by Sparks and collaborators. This discussion focuses on how statistical network modeling and inference can be used to augment the analysis done in their paper. In particular I describe what network models can be used to characterize the dynamics and interactions of Twitter users, and more broadly how network analysis can be used to benefit statistical process monitoring. I hope to not only provide readers a new perspective on how to approach statistical process monitoring in the context of social interactions, but also to motivate future research that address the unique challenges facing quality engineers.

KEYWORDS
Network analysis; statistical process control; statistical process monitoring; syndromic surveillance; Twitter

Introduction

I would first like to congratulate Professor Sparks and collaborators on their development and thorough investigation of strategies for monitoring time between events (TBE) data. The key aim of the presented methodology is to monitor contagion outbreaks among the attendees of the Commonwealth games. To identify outbreaks, the authors develop exponentially weighted moving average plans to monitor the time between the attendees' tweets that contain key phrases related to illness, including, for example, the phrases "coughs," "feeling unwell," and "fever." The hope is that significant increases in tweets about illness signal the onset of an outbreak of some related contagion.

The application of statistical process monitoring (SPM) to syndromic surveillance is a challenging but important endeavor as quality engineers can significantly advance the early detection and management of contagious disease and illness. Although the overall utility of the monitoring plans developed in this paper should certainly be acknowledged, it is my belief that one of the most significant advances in this paper is the authors' use of Twitter data to achieve their goal. Indeed, this application provides a demonstration of the importance and possible power of social media data. Recent news has very clearly established the influence of social media platforms such as Facebook and Twitter—from the dissemination of the #MeToo movement to the motivation of political and industry leaders' actions on women's rights and gun control. The use of social media data arising from these platforms, however, remain largely unexplored by quality engineers and statisticians alike.

Social media data manifest as a collection of measurements over a complicated system describing the demographics, social dynamics, and interactions of users. As a consequence, few have grasped exactly how to make use of such data to enhance their analyses. Making sense of the rich but noisy information from social media platforms is an important but immensely challenging task that needs to be addressed. It is this challenge for which I hope to shed some light in this discussion. I believe that social network analysis is exactly what is needed to effectively incorporate and at least partially make sense of social media platforms, and this opinion is well-supported by past significant analyses of Facebook and Twitter (Ugander et al. 2011; Zaman et al. 2010). Treating the TBE of tweets considered in this paper as a leading example, I will provide simple network analysis strategies—some old and some new—that address two important questions:

1. What network models characterize the dynamics and social interactions of Twitter users?

CONTACT James D. Wilson jdwilson4@usfca.edu Department of Mathematics and Statistics, University of San Francisco, 2130 Fulton Street, San Francisco, CA 94117.
This article was presented at the Sixth Stu Hunter Research Conference in Roanoke, VA, March 2018.

2. How can network analysis benefit SPM strategies like those considered in this paper?

For the first question, I will discuss two families of well-established network models that readily fit the problem at hand—*naturally occurring networks*, and *probabilistic graphical models*. For the second question, I propose a few simple network analysis strategies that I believe will provide significant insights to the application considered. In this discussion I hope to not only provide the readers a new perspective on how to approach SPM in the context of social interactions, but also to motivate future research that address the unique challenges facing quality engineers.

Why use networks for statistical process monitoring?

The study of networks has been motivated by the modeling and understanding of complex systems. Networks are used to model the relational structure between individual units of an observed system. Network-based models have been used in a variety of disciplines: in biology to model protein-protein and gene-gene interactions; in sociology to model friendship and information flow among a group of individuals; and in neuroscience to model the relationship between the organization and function of the brain.

Network analysis stems, and has had a profound impact in, the social sciences where questions center around the dynamics of individuals. From Moreno's first use of a social network in Moreno and Jennings (1934) to modern day analyses of social media (Ugander et al. 2011; Bhamidi et al. 2015), our understanding of social interactions has greatly improved (see Wasserman and Faust (1994) for an extensive treatment on the topic). Statistical and computational advances have further enabled the modeling and analysis of large data sets like those arising from social media (Goldenberg et al. 2009).

One can leverage the rich literature of social network analysis to enhance the current capabilities of SPM especially in the presence of social media data. In particular, social network analysis provides strategies that can be used to directly analyze relational data and should be incorporated in SPM analysis of social media for (at least) the following two reasons: (i) network models provide a richer understanding of social media users than demographic and TBE measurements alone, and (ii) network analysis enables the monitoring of both global (considered in this paper) and local signals in the system under surveillance.

I claim that the first reason is self-evident based on the tremendous development and application of network analysis techniques over the past four decades. Having said that, the choice of *which* network model still requires careful consideration and reasonable knowledge and exploration of the data being studied. It also requires an understanding of which network models provide meaningful representations of the data and what subsequent analyses of the network could possibly discover. To provide some intuition in the case of Twitter data, I will provide three different network models for the monitoring of TBE on Twitter.

In the paper being discussed, the authors monitor the TBE, $\{w_1, \ldots, w_{n-1}\}$, of illness-related tweets among an entire population of Twitter users. By monitoring the entire population, however, only *global* outbreaks, or outbreaks that occur across the population, can be detected. Thus, this strategy does not account for possible *localized* outbreaks—contagion outbreaks that occur among a smaller group of users, whose group perhaps contains users from a similar geographic location, or users who attended the same concert event or boarded the same airline from which a contagion originated. Armed with the networks describing the Twitter users under study, one can readily monitor localized outbreaks in addition to the global outbreaks of the original monitoring plan. I discuss this more fully below.

Networks describing twitter and time between events data

The first task in any network analysis is to determine what network models are appropriate for the observed data and the question at hand. That is, one needs to construct a network model $G = (V, E)$ so that the vertex set V represents the actors or individuals of interest, and the edge weights E quantify the strength of dependence between pairs of actors. In the case of Twitter data, V almost always represents the users on Twitter. The choice of E, on the other hand, requires more thought. In this application, one can construct at least three models for E without much effort, each of which provide different and potentially useful information about the relationships of the users. Next, I will describe these three models—two arising directly from the social relationships of the users, and the other as a probabilistic graphical model that describes the dependence between users as measured from TBE data.

Naturally occurring networks for twitter: Follower and re-tweet networks

Perhaps the most common two networks used to analyze Twitter are the Follower and retweet networks (see, e.g., Bhamidi et al. (2015)). The Follower and retweet habits of users on Twitter are each example of what is sometimes referred to as a *naturally occurring network*. A naturally occurring network on actors V exists if there are any relational measurements taken on the actors, namely measurements that are taken over pairs of actors. For the Follower network, one observes the following pairwise measurements for all $u, v \in V$:

$$x_F(u,v) = \mathbb{I}(u \text{ follows v}).$$

For the retweet network, the following binary measurements are observed for each pair of actors $u, v \in V$:

$$x_{RT}(u,v) = \mathbb{I}(u \text{ hasre-tweeted v during the data collection process}).$$

Notably, the quantity $x_{RT}(u,v)$ could also be specified to quantify the number of times u retweets v. One also needs to be careful about the length of time over which these values are measured. The subsequent analysis would rely on techniques appropriate for weighted networks (Wilson et al. 2017). Whatever the choice, the Follower and retweet networks are the directed networks $G = (V, E)$ with edge weights $E = \{x_F(u,v) : u, v \in V\}$ and $E = \{x_{RT}(u,v) : u, v \in V\}$, respectively.

Probabilistic graphical models for time between events

Professor Sparks and collaborators monitored the TBE, $\{w_1, \dots, w_{n-1}\}$, of an illness-related tweet over the entire population. It is likely, however, that each user's tweet is dependent upon the tweets of other users in the population. If the social structure of the Twitter users or social media data under consideration is not available, it is still possible to construct a network describing the users' relationships using individual TBE using probabilistic graphical models.

Undirected graphical models, also known as Markov networks, have a long history and are now ubiquitous in statistical machine learning (see Koller and Friedman (2009); Wainwright and Jordan (2007) for book-level treatments of the topic.) Given a vector of random variables $X = [X_1, \dots, X_p]$, an undirected graphical model for X is the graph $G = (V, E)$ with vertex set $V = \{1, \dots, p\}$ and edge set E containing pairs (u, v) for which X_u is conditionally dependent upon X_v given the remaining variables $\{X_j : j \neq u, v\}$. By construction, the graph G represents a first order Markov dependence between the variables of X.

Much of the research on undirected graphical models has focused on the family of Gaussian graphical models, under which X is assumed to be a multivariate Gaussian random vector. Yang et al. (2015) very recently extended the foundations of Gaussian graphical models to random vectors from multivariate exponential families. It is that work that enables the estimation of a probabilistic graphical model for the TBE data for Twitter users. Let $w_t^{(u)}$ denote the time between the tth and $t+1$st event for user u. Set $W_t := [w_t^{(u)} : u \in V]$ to be the vector of these TBE for each user. Under the assumption that W_t is a random vector from some multivariate exponential family, it is possible to estimate a graph at time t using the M-estimation strategy described in Yang et al. (2015). If the TBE is assumed to a multivariate exponential random vector, an example explored in Professor Sparks' paper, one can estimate the graphical model characterizing the TBE vector W_t. The precise details of this model is provided in equation (15) of Yang et al. (2015).

The above strategy presents just one example of a probabilistic graphical model for TBE, though others are possible. Future work should investigate how to estimate an exponential model with temporal dependence as well as for other distributions like the Gamma distribution described in Sparks' work.

Monitoring TBE using social networks data

Once a network model (or some collection of models) has been chosen, one can augment the monitoring strategy on TBE using network characteristics. The monitoring of networked data has recently gained a lot of attention, but new methods are needed (see Jeske et al. (2018) and Woodall et al. (2017) for recent reviews). For the discussion of TBE, I will revisit the challenge of monitoring the system for local outbreaks in addition to global ones. I propose three subgraph-based strategies to provide some intuition as to what is possible. These strategies are motivated by the homophily principle (McPherson et al. 2001), which posits that vertices with similar external characteristics are highly connected to one another in the network.

1. **Neighborhood TBE**: In unweighted undirected networks the neighborhood of vertex u, $\mathrm{Ne}(u)$, is defined as the collection of vertices that share an edge with u in V. Analogous definitions are

available for directed and weighted graphs, but I omit them here. Local outbreaks can be detected through the monitoring of the TBE for vertices in each neighborhood of G. That is, the *neighborhood TBE* given by $\mathbf{w}_{\mathrm{Ne}(u)} = \{w_i^{(v)} : v \in \mathrm{Ne}(u)\}$ can be monitored for each node u.

2. **Clique TBE**: A clique is a complete subgraph of vertices, namely a collection of vertices where every pair of vertices contains an edge between them. For each vertex u, let $\mathrm{Cl}(u)$ denote the largest clique for which u belongs. Note that the vertices in the clique of u is a more strongly connected subset of the vertices belonging to the neighborhood of u and hence represents a collection of vertices that demonstrate strong clustering. Once the maximal clique for each vertex has been identified, the *clique TBE* given by $\mathbf{w}_{\mathrm{Cl}(u)} = \{w_i^{(v)} : v \in \mathrm{Cl}(u)\}$ can be monitored.
3. **Community TBE**: Empirically the nodes of a network G can often be divided into $k \geq 1$ disjoint vertex sets as $V = V_1 \cup V_2 \ldots \cup V_k$ in such a way that the density of edges within each vertex set $V_j \subseteq V$ is substantially greater than the density between differing sets. These densely connected vertex sets are commonly referred to as *communities*. In many applications, the communities of a network provide structural or functional insights about the modeled complex system. For example, recently community structure has been used to help develop hypotheses about gene interactions and antibiotic resistance (Parker et al. 2015), about the dynamics of social interactions using cell phone data (Greene et al. 2010), and in identifying functional subregions of the brain (Stillman et al. 2017). The substantial relevance of communities in network systems has led to a large and growing literature about community structure and the identification of statistically meaningful communities (Wilson et al. 2014; Porter et al. 2009; Fortunato 2010). With the communities in hand, one can monitor the *community TBE* $\mathbf{w}_j = \{w_i^{(v)} : v \in V_j\}$ for each community $j = 1, \ldots, k$.

The above three strategies provide a straightforward manner to augment the analysis of TBE. It should be noted that these strategies generalize to any statistics measured on the individuals, including for example the counts of events considered in this paper. An example of a monitoring plan that investigates local and global network changes is described in (Sparks and Wilson 2016). Further, Wilson et al. (2016) investigated the monitoring of networks with community structure that change through time. Higher order subgraph structures, like triads or cycles can also be investigated. Finally, changes in the overall generative process describing the observed network through time can also be monitored. Such analyses rely upon the appropriate definitions of dynamic random graph models that characterize the temporal dependence between networks. There are several works in this area to consider, including dynamic versions of the exponential random graph model (Hanneke et al. 2010; Krivitsky and Handcock 2014; Lee et al. 2017), latent space networks (Sewell and Chen 2015), as well as stochastic block models (Wilson et al. 2016; Xu and Hero 2014).

As discussed earlier, there are multiple network models that describe the Twitter users in this study. Together, these different models form a multiplex network model of the Twitter users. To utilize the information from each of the network representations, multiplex network methods can be used (see Kivelä et al. (2014) for a recent review). To provide a concrete example, for the *community TBE* defined above, communities can be identified using multilayer network community detection methods like those available in Mucha et al. (2010), De Domenico et al. (2015), and Wilson et al. (2017).

Concluding remarks

I would like to thank the organizers of the Sixth Stu Hunter conference for giving me the opportunity to discuss this work. Professor Sparks and collaborators have set the stage in the development of SPM methodology to social media data for syndromic surveillance, yet important challenges still face the quality engineering community. I have discussed and sought to address one major challenge, which is how to utilize social media data to enhance the application of SPM. My discussion focused on the use of network analysis for social media platforms. I described strategies for how to choose an appropriate network model for the dynamics and interactions of individuals, as well as how to subsequently utilize these networks to monitor events from the individuals. I hope that this discussion provides a new lens from which quality engineers can view the problem of SPM. Moreover, I hope that the proposed strategies here motivate future analyses that address the unique challenges of monitoring networked data. I look forward to what is to come.

About the author

James D. Wilson is an Assistant Professor of Statistics and Data Science at the University of San Francisco. He is also the Co-Director of Data Science and Associate Director of Research of the Data Institute at the University of San Francisco. He received his Ph.D. of Statistics and Operations Research at the University of North Carolina at Chapel Hill in 2015. His research brings together techniques from machine learning, statistical inference, and random graph theory to model, analyze, and explore relational (network) data. He is particularly interested in developing random graph models and feature extraction methodologies for dynamic and multilayer networks; monitoring networked systems; and investigating networks that arise in diverse applications ranging from neuroscience to political science to infectious disease. He is a member of the ASA, ACM, and IMS.

ORCID

James D. Wilson http://orcid.org/0000-0002-2354-935X

References

Bhamidi, S., J. M. Steele, and T. Zaman. 2015. Twitter event networks and the superstar model. *The Annals of Applied Probability* 25 (5):2462–502.

De Domenico, M., A. Lancichinetti, A. Arenas, and M. Rosvall. 2015. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Physical Review X* 5 (1):011027.

Fortunato, S. 2010. Community detection in graphs. *Physics Reports* 486 (3–5):75–174.

Goldenberg, A., A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. 2009. A survey of statistical network models. *Foundations and Trends® in Machine Learning* 2 (2): 129–233.

Greene, D., D. Doyle, and P. Cunningham. 2010. Tracking the evolution of communities in dynamic social networks. Paper presented at 2010 International conference on advances in social networks analysis and mining (ASONAM), 176–183, Odense, Denmark, August 9.

Hanneke, S., W. Fu, and E. P. Xing. 2010. Discrete temporal models of social networks. *Electronic Journal of Statistics* 4:585–605.

Jeske, D. R., N. T. Stevens, A. G. Tartakovsky, and J. D. Wilson. 2018. Statistical methods for network surveillance. *Applied Stochastic Models in Business and Industry* 34(4). doi:10.1002/asmb.2326.

Kivelä, M., A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. 2014. Multilayer networks. *Journal of Complex Networks* 2 (3):203–71.

Koller, D., and N. Friedman. 2009. *Probabilistic graphical models: principles and techniques*. Cambridge, MA: MIT Press.

Krivitsky, P. N. and M. S. Handcock. 2014. A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76 (1): 29–46.

Lee, J., G. Li, and J. D. Wilson. 2017. Varying-coefficient models for dynamic networks. *Preprint arXiv*:1702: 03632.

McPherson, M., L. Smith-Lovin, and J. M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27 (1):415–44.

Moreno, J. L. and H. H. Jennings. 1934. *Who shall survive*? Washington, DC: Nervous and Mental Disease Publishing.

Mucha, P. J., T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela. 2010. Community structure in time-dependent, multiscale, and multiplex networks. *Science* 328 (5980):876–8.

Parker, K. S., J. D. Wilson, J. Marschall, P. J. Mucha, and J. P. Henderson. 2015. Network analysis reveals sex-and antibiotic resistance-associated antivirulence targets in clinical uropathogens. *ACS Infectious Diseases* 1 (11): 523–32.

Porter, M. A., J.-P. Onnela, and P. J. Mucha. 2009. Communities in networks. *Notices of the AMS* 56 (9): 1082–97.

Sewell, D. K., and Y. Chen. 2015. Latent space models for dynamic networks. *Journal of the American Statistical Association* 110 (512):1646–57.

Sparks, R. and J. D. Wilson. 2016. Monitoring communication outbreaks among an unknown team of actors in dynamic networks. *Preprint arXiv*:1606:09308.

Stillman, P. E., J. D. Wilson, M. J. Denny, B. A. Desmarais, S. Bhamidi, S. J. Cranmer, and Z.-L. Lu. 2017. Statistical modeling of the default mode brain network reveals a segregated highway structure. *Scientific Reports* 7 (1): 11694.

Ugander, J., B. Karrer, L. Backstrom, and C. Marlow. 2011. The anatomy of the facebook social graph. *Preprint arXiv*:1111:4503.

Wainwright, M. J. and M. I. Jordan. 2007. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1 (1–2): 1–305.

Wasserman, S. and K. Faust. 1994. *Social network analysis: Methods and applications*, vol.8. New York, NY: Cambridge University Press.

Wilson, J. D., M. J. Denny, S. Bhamidi, S. J. Cranmer, and B. A. Desmarais. 2017. *Stochastic weighted graphs: Flexible model specification and simulation. Social Networks* 49, 37–47.

Wilson, J. D., J. Palowitch, S. Bhamidi, and A. B. Nobel. 2017. Community extraction in multilayer networks with heterogeneous community structure. *The Journal of Machine Learning Research* 18 (1):5458–506.

Wilson, J. D., N. T. Stevens, and W. H. Woodall. 2016. Modeling and estimating change in temporal networks via a dynamic degree corrected stochastic block model. *Preprint arXiv*:1605:04049.

Wilson, J. D., S. Wang, P. J. Mucha, S. Bhamidi, and A. B. Nobel. 2014. A testing based extraction algorithm for identifying significant communities in networks. *The Annals of Applied Statistics* 8 (3):1853–91.

Woodall, W. H., M. J. Zhao, K. Paynabar, R. Sparks, and J. D. Wilson. 2017. An overview and perspective on social network monitoring. *IISE Transactions* 49 (3): 354–65.

Xu, K. S. and A. O. Hero. 2014. Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing* 8 (4): 552–62.

Yang, E., P. Ravikumar, G. I. Allen, and Z. Liu. 2015. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research* 16 (1):3813–47.

Zaman, T.R., R. Herbrich, J. Van Gael, and D. Stern. 2010. Predicting information spreading in twitter. Paper presented at Workshop on computational social science and the wisdom of crowds, NIPS, vol. 104, 17599–601, Vancouver, Canada, December 10.