Evidence of Temporal Artifacts in Social Networks

Matt Revelle, Carlotta Domeniconi, and Aditya Johri

George Mason University,
Fairfax VA 22030, USA
{revelle,carlotta}@cs.gmu.edu,ajohri3@gmu.edu

Abstract. There has been extensive research on social networks and methods for specific tasks such as: community detection, link prediction, and tracing information cascades; and a recent emphasis on using temporal dynamics of social networks to improve method performance. The underlying models are based on structural properties of the network, some of which we believe to be artifacts introduced from common misrepresentations of social networks. Specifically, representing a social network or series of social networks as an accumulation of network snapshots is problematic. In this paper, we use a dataset with timestamped interactions to demonstrate how cumulative graphs differ from activity-based graphs and may introduce temporal artifacts.

1 Introduction

The modeling of social networks is an expansive and active area of research. While models may incorporate other network features such as node attributes [4, 24, 16], nearly all rely on network structure. Many methods are now also incorporating temporal dynamics [12, 10, 20, 22], but how the temporal information is integrated varies. There are various approaches [20, 21] to representing a dynamic social network as a series of networks, but until recently [15] all have lacked theoretical foundation.

Dynamic network representations which capture edge deactivation [20] have shown to improve task-specific performance. However, many state-of-the-art methods [23, 16] are based on *cumulative graphs* and ignore edge deactivation. The findings presented in this paper suggest that some existing models may be designed to accomodate temporal artifacts introduced by not including edge deactivation in the processing of network data.

There are two social network phenomena which motivate our analysis: *social capacity* [3] and *bursty events* [1]. Social capacity can be viewed as a per-node limit on the number of incident edges active at any given time and thus conflicts with the claim of densification and shrinking diameters in social networks [8, 11] unless additional conditions are met. For example, a network where every new node has a larger social capacity would lead to densification and shrinking diameters. While variation in social capacity based on demographics has been observed [15] there has been no evidence presented that would indicate social capacity is a function of when a node joins the network.

In order to measure the existence of densification and shrinking diameters, we first must construct a series of network snapshots which more accurately captures network structure than simply accumulating all edges over time. We do this by using communication activity between nodes as evidence that an edge is active. The bursty dynamics of social communication are accounted for by measuring the inter-event times and selecting an observation window large enough to minimize incorrectly deactivating an active edge. Thus we are able to construct a series of *activity graphs* which provide a more accurate approximation of the network state at a given point in time. This method of graph construction has been used previously on a mobile phone network [15] to improve understanding of communication strategies. We can then measure and compare evidence of densification and shrinking diameters in both a cumulative graph series and an activity graph series.

Densification and diameter shrinking are accepted as basic characteristics of dynamic social networks. However, this paper presents results which contradict those findings. When edge deactivation is incorporated, we do not find evidence of densification and diameter shrinking. We suggest this may be an effect of social capacity.

2 Related Work

Existing methods for social network tasks have either ignored temporal dynamics [16, 24] or proposed methods to filter edges with a decay function [20] or sliding window [21]. While these attempts to account for temporal dynamics may be effective, they are ad-hoc and lack a theoretical justification. The work by Miritello et al. [14] proposes the selection of an observation window size based on inter-event statistics and a simple method for identifying edge activation and deactivation. While similar to existing sliding window approaches, this method is motivated by social interaction patterns (bursty events). This approach is used to construct the activity graph series for our experiments.

Models of dynamic social networks based on node interaction activity [9, 17] have been introduced. These models are capable of generating single network snapshots which resemble real world networks. The existing models are unable to produce a graph series which corresponds to a real-world network series. However, their ability to generate networks with realistic structure indicates they are an alternative to previous models which heavily rely on preferential attachment [2] or community affiliation [23] and ignore social interaction patterns. There are many types of temporal networks [6] and this paper presents observations on dynamic social networks, specifically person-to-person communication networks.

3 Background

The concepts of social capacity [3] and bursty communications [19, 1] have been considered separately and recent literature [15, 13, 14] has attempted to measure and use these to determine the state of edges in a large social network.

Social capacity captures the maximum number of relationships one prefers to maintain at any given time and there is evidence that social capacity is conserved over time [15, 5, 7]. The term bursty is used to describe the temporal patterns of social interactions between pairs of nodes. That is, humans tend to interact in bursts and these patterns must be considered in order to correctly identify the activation/deactivation of edges.

The observation of social capacity and burstiness of human interaction in some networks suggests careful consideration is required to construct accurate static views of these networks. In fact, accepted claims of graph evolution [8, 11] appear to fail when graph series are constructed based on timestamped interactions rather than accumulated without regard for edge deactivation.

Previous literature [23] introduced densification and diameter shrinking as common network characteristics and we briefly describe them here. Densification is the super-linear growth of edges relative to nodes and results in a network becoming denser over time. Diameter shrinking is the reported tendency for network diameters to decrease over time as more edges are accumulated. We can see both how densification may lead to diameter shrinking and contradicts the notion of social capacity.

4 Evidence of Temporal Artifacts

4.1 Dataset Descriptions

A dataset with timestamped interactions is required to construct an accurate temporal series of networks. We use data from Scratch [18], an online community where users may write and share projects (programs). There are several ways by which Scratch users may interact: project comments, project remixes, gallery curation, and user following. More information about Scratch may be found in [18]. We selected a single type of interaction to simplify analysis. Project comments are a natural choice as they are the most-frequent interaction between Scratch users and thus a better approximation of edge status (active/inactive). These project comments serve as a means for users to communicate within the context of a project. The comments in the Scratch dataset are timestamped and thus we can create timestamped edges from comment authors to the project authors.

The dataset spans over March 2007 to December 2011 and includes a large period of rapid growth in Scratch users, shown in Figure 1, which does not slow until towards the end of the dataset. There are a total of 7,788,000 interactions between 164,205 users. We use all these interactions when constructing the graph series. However, there are many short-term interactions and we filter out directed interactions between pairs which only occur once or twice when measuring communication behavior. Such interactions have undefined or trivial inter-event statistics as there are zero or one inter-event observations when only one or two interactions are observed. There are a total of 1,799,050 of such interactions with frequency < 3 which were filtered, leaving 5,988,950 interactions. The Scratch dataset used to construct the networks may be obtained from the MIT Media Lab website[1].

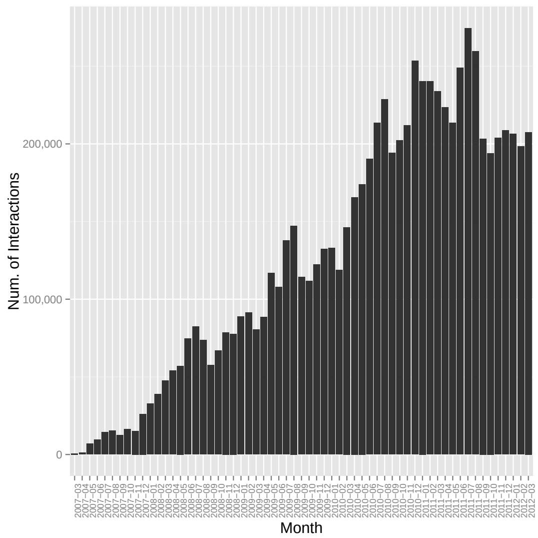

Fig. 1: The number of interaction events occurring by month.

4.2 Methodology

As the relationships in the Scratch interaction network are based on communication events between nodes, we check for evidence of bursty patterns. Bursty communication can be identified by the dispersion of inter-event times between node pairs. If communication is bursty then the standard deviation of inter-event time will be larger than the mean. The ratio of the mean and standard deviation of inter-event times is the coefficient of variation (cv) and used to measure dispersion. When $cv > 1$, there is evidence of bursty communication. The use of dispersion to identify burstiness is further discussed by Miritello et al.[14].

We hypothesize the observation of densification and diameter shrinking [8, 11] may be attributed to the inclusion of deactivated edges in a network. To test this we construct two graph series. The series are both constructed from the Scratch dataset and each network in the series captures network activity over consecutive and non-overlapping three-month periods. The three-month length of the observation window was selected because it is large enough to account

[1] https://llk.media.mit.edu/scratch-data

for the majority of inter-event times (97% of inter-event times are < 62 days) and conveniently maps to annual quarters. The first series is a *cumulative graph series* where new nodes and edges are added at each consecutive snapshot to the previous network in the series. The second series is based on node interaction activity and we refer to it as the *activity graph series.*

Edge activity is determined by tracking the activation and deactivation of edges based on observations in a three-month window along with the previous and next three-month periods. A similar approach has been used in previous literature [14]. An edge is considered to activate if it is not present in the three months preceding the three-month observation window but an event falls within the observation window. Similarly, an edge is deactivated if an event occurs in the observation window but not in succeeding three months. Only edges active in each three-month observation window are used in the corresponding graph in the activity graph series.

The edge-node ratio ($\frac{num.of edges}{num.of nodes}$) is calculated for each graph in both series and used to measure densification. If densification is present, we expect the number of edges to grow super-linearly in the number of nodes [11]. We also measure the diameter of every graph in both series to determine whether diameter shrinking is observed.

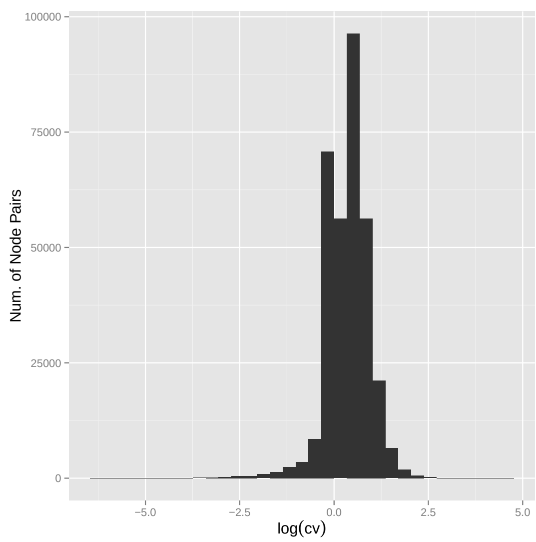

Fig. 2: The $log(cv)$ for node pairs in the Scratch interaction network with $>= 3$ events. A small number of node pairs (1,038) were removed for this plot as they had a cv of zero and thus were undefined.

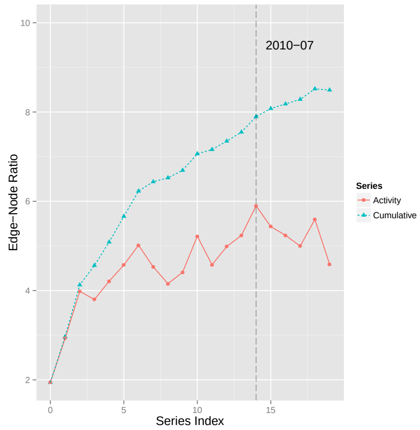

Fig. 3: The edge-node ratio over time in the cumulative and activity graph series.

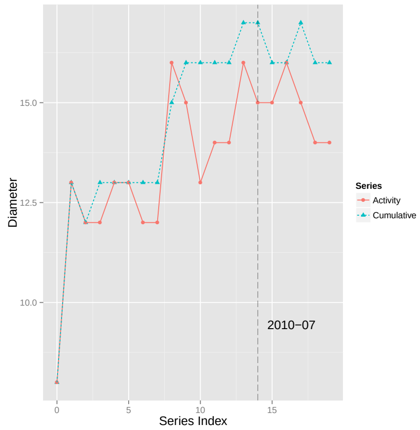

Fig. 4: The diameter length over time in the cumulative and activity graph series.

4.3 Results

As shown in Figure 2, bursty communication patterns are observed as the cv values are frequently > 1 ($log(cv) > 0$).

We see evidence of densification in the cumulative series but not in the activity series, in Figure 3. The accumulation of edges, without removal of deactivated edges, appears to introduce densification as a temporal artifact in the Scratch interaction network. This is especially clear when the number of interactions stops growing around July 2010, denoted by dashed vertical line in both Figures 3 and 4.

Surprisingly, an overall trend of diameter shrinking is not clearly observed in either network series. Figure 4 shows a generally increasing diameter for both series and a larger variance in diameter for the activity series. The lack of diameter shrinking may be due to the growth of the Scratch website during most of this time period. Both include a vertical line marking the month (July 2010) when the increase in the number of Scratch interactions slows.

These findings are not unexpected but they are contrary to previous literature [8, 11] which has served as the basis for state-of-the-art network models. The edge-node ratio in the cumulative graphs is monotonically increasing over time and social capacity is ignored. In contrast, the edge-node ratio in activity graphs may decrease or stabilize as inactive edges are detected and removed.

5 Conclusion

This paper presents evidence that temporal artifacts may be introduced in social networks when the relationships represented by edges require allocation of inelastic resource such as time or attention. Our findings suggest more accurate social networks may be derived from ongoing dyadic interactions rather than one-time events such as "following" or "friending."

We plan to extend this work to include other datasets, explore how community affiliation correlates to interaction patterns, and ultimately provide a model of social networks which incorporates knowledge from these findings.

6 Acknowledgment

We appreciate the Lifelong Kindergarten group at MIT for publicly sharing the Scratch datasets. This work is partly based upon research supported by U.S. National Science Foundation (NSF) Awards DUE-1444277 and EEC-1408674. Any opinions, recommendations, findings, or conclusions expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

References

1. A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.
2. A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3):590–614, 2002.
3. B. Gonçalves, N. Perra, and A. Vespignani. Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS one*, 6(8):e22656, 2011.
4. S. Günnemann, B. Boden, I. Färber, and T. Seidl. Efficient mining of combined subspace and subgraph clusters in graphs with feature vectors. In *Advances in Knowledge Discovery and Data Mining*, pages 261–275. Springer, 2013.
5. C. A. Hidalgo and C. Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024, 2008.
6. P. Holme and J. Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.
7. G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.
8. R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*, pages 337–357. Springer, 2010.
9. G. Laurent, J. Saramäki, and M. Karsai. From calls to communities: a model for time varying social networks. *arXiv preprint arXiv:1506.00393*, 2015.
10. J. Leskovec. Social media analytics: tracking, modeling and predicting the flow of information through networks. In *Proceedings of the 20th international conference companion on World wide web*, pages 277–278. ACM, 2011.
11. J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.
12. Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 6–14. ACM, 2012.

13. G. Miritello, R. Lara, M. Cebrian, and E. Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3, 2013.
14. G. Miritello, R. Lara, and E. Moro. Time allocation in social networks: correlation between social structure and human communication dynamics. In *Temporal Networks*, pages 175–190. Springer, 2013.
15. G. Miritello, E. Moro, R. Lara, R. Martínez-López, J. Belchamber, S. G. Roberts, and R. I. Dunbar. Time as a limited resource: Communication strategy in mobile phone networks. *Social Networks*, 35(1):89–95, 2013.
16. F. Moser, R. Colak, A. Rafiey, and M. Ester. Mining cohesive patterns from graphs with feature vectors. In *Proceedings of the SIAM International Conference on Data Mining*, volume 9, pages 593–604. SIAM, 2009.
17. N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani. Activity driven modeling of time varying networks. *Scientific reports*, 2, 2012.
18. M. Resnick, J. Maloney, A. Monroy-Hernández, N. Rusk, E. Eastmond, K. Brennan, A. Millner, E. Rosenbaum, J. Silver, B. Silverman, et al. Scratch: Programming for all. *Communications of the ACM*, 52(11):60–67, 2009.
19. M. T. Rivera, S. B. Soderstrom, and B. Uzzi. Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *annual Review of Sociology*, 36:91–115, 2010.
20. R. Rossi and J. Neville. Modeling the evolution of discussion topics and communication to improve relational classification. In *Proceedings of the First Workshop on Social Media Analytics*, pages 89–97. ACM, 2010.
21. R. A. Rossi, B. Gallagher, J. Neville, and K. Henderson. Modeling dynamic behavior in large evolving graphs. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 667–676. ACM, 2013.
22. Y. Sun, J. Tang, J. Han, M. Gupta, and B. Zhao. Community evolution detection in dynamic heterogeneous information networks. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 137–146. ACM, 2010.
23. J. Yang and J. Leskovec. Community-affiliation graph model for overlapping network community detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1170–1175. IEEE, 2012.
24. J. Yang, J. McAuley, and J. Leskovec. Community detection in networks with node attributes. In *IEEE 13th International Conference on Data Mining*, pages 1151–1156. IEEE, 2013.