Thomas Magelinski
tmagelin@andrew.cmu.edu

Prof. Kathleen M. Carley
Kathleen.carley@cs.cmu.edu

Carnegie Mellon University
www.casos.cs.cmu.edu

Effects of Network Aggregation in Simple Diffusion Simulations

Abstract

Diffusion models are at the heart of many pressing issues such as disease, fake news, and computer viruses. Epidemiologists and others have recognized the importance of link structure in static networks for diffusion simulations. At the same time, network scientists have recognized the importance of temporal structure in complex systems. In this work, we bring the two together to study the effect of condensing a temporal network into a static network in the context of diffusion. Using the College Message Dataset and a simple Susceptible-Infected simulation model, we study the differences in diffusion given three networks: the original temporal network, the static binarized network, and the static thresholded network. **We find that the static networks exhibit significantly different diffusion properties to the original temporal network**. In general, diffusion occurred more quickly and reached more of the nodes given the binarized network as opposed to the temporal network. The opposite was true for the tresholded network. Additionally, the choice of seed node lead to drastically different results in the temporal network, while there was no difference in the static networks. Lastly, we found that low transmissibility simulations on the temporal network further increased the disparity between seed nodes, showing promise as a method of finding temporally-central noes. These findings are in line with the theory of network cascades.

Dynamic Networks and Aggregation

We consider three types of networks which are typically used in practice:

1. **Temporal Networks**: All links have a start time and an end time. These networks are represented as *network snapshots* or series of static networks. Temporal networks fully capture the dynamics of the data.
2. **Binarized Networks:** A single static network. Links are binary and indicate if that connection took place anywhere in the timeline.
3. **Thresholded Networks:** A single static network. Links are first weighted by the number of occurrences in the timeline. Then, weights less than a threshold (the mean link weight for our study) are removed. Finally, all remaining link weights are set to one.

A simple example of these networks is shown in Figure 1.

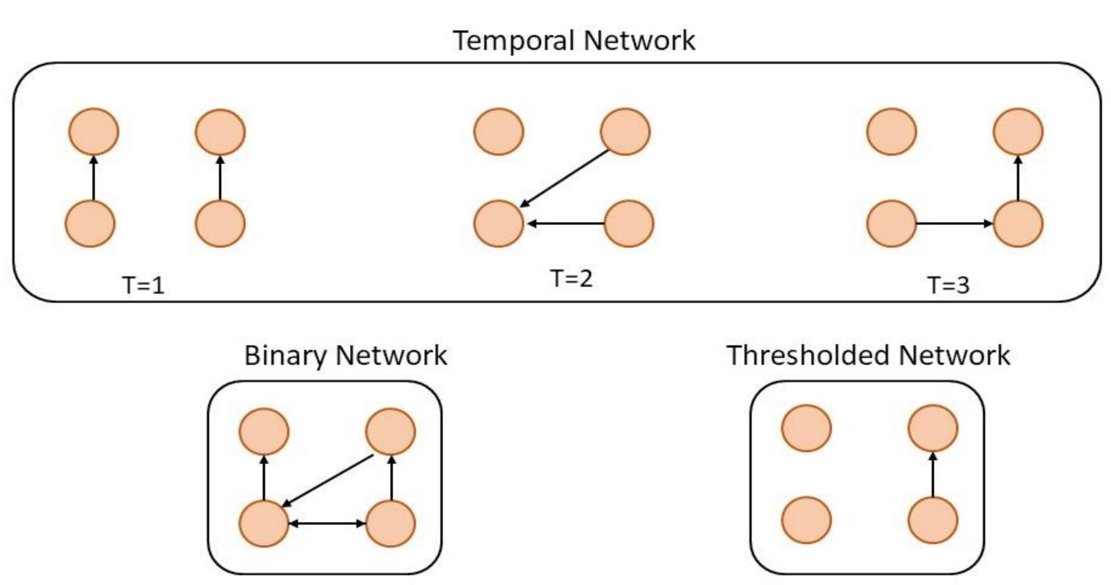

Figure 1: Three types of networks analyzed in this study.

Figure 1 is a very simple network, but aggregation already harms the data. In the binary network, diffusion can occur from the top right to the bottom left, then to the top left. However, this is not possible in the original temporal network due to the order of the edges. In the thresholded network, most of our data is lost in the thresholding procedure. We aim to understand how these effects playout on a large, real social media dataset.

Results

Both static networks showed very different diffusion patterns to the original temporal network. Diffusion occurred faster on the binary network since all of the links were available to diffuse over at every time step. The thresholded network was hardest to diffuse on, since it had the fewest links. The seed node mattered most in the temporal network. All differences were magnified as p decreased. For every seed, **both the type of network and the transmissibility was significant at the p <0.05 level**. Interaction between the two variables was not found to be significant.

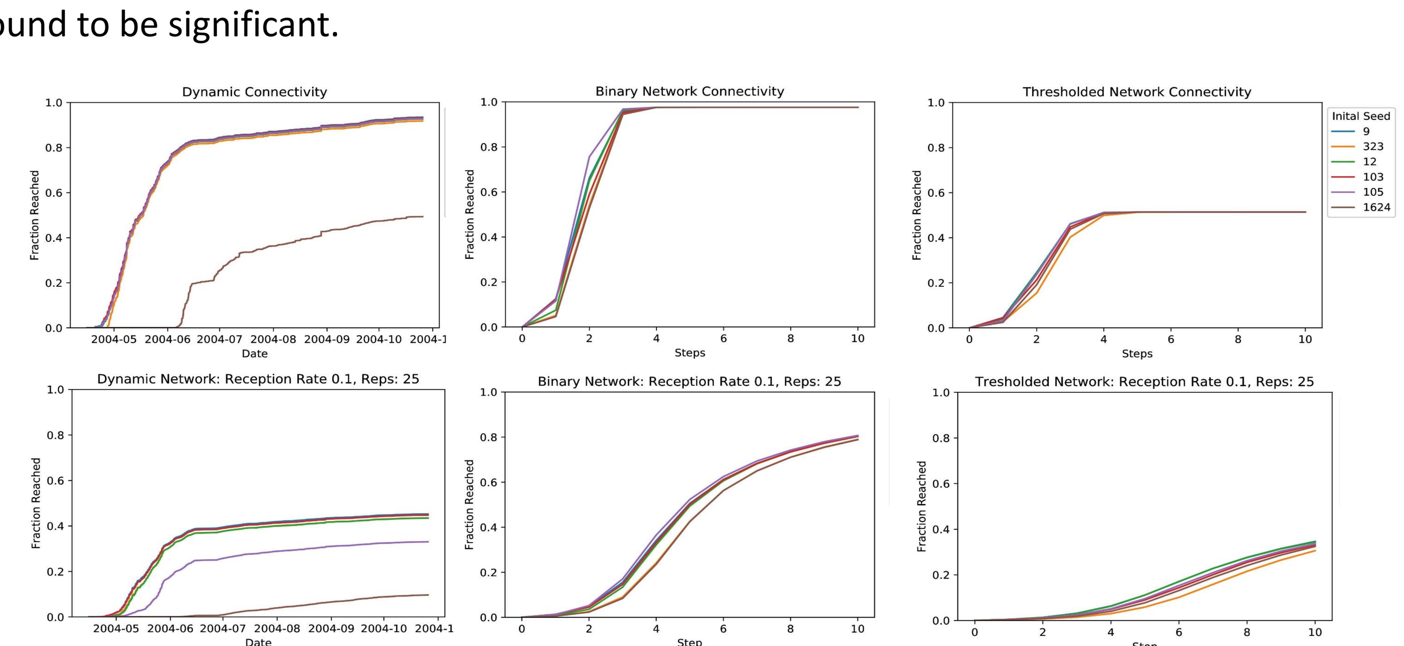

Figure 2: Average results for all 3 networks for p=1 and p=0.1.

The College Message Dataset

In this study we analyze the College Messaging Dataset*. The dataset was constructed by collecting private messages sent on an online social media platform at the University of California, Irvine, The dataset consists of source nodes (message senders), target nodes (message receivers), and time-stamps.

Table 1: Basic statistics of the College Message Dataset

Feature	Value
Nodes	1,899
Messages	59,835
Time-span	193 Days

*Panzarasa, P., Opsahl, T., Carley, K.M.: Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. Journal of the American Society for Information Science and Technology 60 (5) , 911–932 (2009).

Experimental Setup

Diffusion over the college message dataset was modeled with a simple Susceptible-Infected (SI) model. This model captures diffusion of knowledge, where students either have heard the fact (infected) or have not (susceptible). Infected nodes try to spread the knowledge through their links. They are successful with transmission probability p, the only parameter in the model. **In the temporal network, simulation occurs on the real time line**, since agents try to spread knowledge over messages in real time. **In the static networks, simulation occurs on abstract time steps**, since nodes can only reach their neighbors, the process is repeated iteratively to model the concept of time.

The model also requires a seed node and a seed time. This dictates who starts with the idea and when. In all of our experiments only a single student starts with the idea, and they have that idea at time t=0. The actual seed varies between the top-6 students in terms of messages sent.

Table 2 summarizes the experimental setup. Each experiment was repeated 25 times, for a total of 918 total runs. In all cases, the outcome variable was the fraction of infected nodes over time. The time-series were compared qualitatively while the final fraction infected across experiments were compared quantitatively using a two-way ANOVA.

Table 2: Experimental Design

Variable	Cases	Implication
Network Type	Temporal, Binary, Thresholded	Diffusion environment
Infection Rate	1, 0.5, 0.1	Likelihood of spreading idea
Initial Seed	9, 323, 12, 103, 105, 1624	Who starts with the idea
Seed Time	0	When the seed gets the idea
Number of Seeds	1	How many seeds there are

Conclusions

Using the College Message Dataset and a simple Susceptible-Infected simulation model, we see that **static networks exhibit drastically different diffusion properties then the underlying temporal network**. In general, diffusion occurred more quickly and reached more of the nodes in the binarized network than in the temporal network.

This highlights the importance of retaining temporal structure of network data when possible. We hope that this work will be used as a stepping stone for future work guiding practitioners for best practices in simulation in the absence of sufficient temporal granularity.

This material is based upon work supported by the Office of Naval Research Multidisciplinary University Research Initiative (MURI) under award number N00014-17-1-2675. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research. Additionally, Thomas Magelinski was supported by an ARCS foundation scholarship.

institute for SOFTWARE RESEARCH