Chapter 5

Conclusion

In this thesis, we developed two new techniques for exploratory data analysis in dynamic networks, and exposed inherent measurement biases in a third. The focus has been on analyzing massive dynamic network datasets to understand facets of the physical systems that they represent. As researchers find more ways to record temporal information about network datasets, often on a continuous streaming basis, the need for exploratory tools that are agnostic to the source of data becomes apparent. These tools must make minimal assumptions about the underlying system, and be computationally and statistically tractable. The two new tools we develop fall into this category: the Fourier-like decomposition for dynamic networks in Chapter 2 assumes only the existence of locally periodic behavior, and the coupled edges we develop in Chapter 3 extend the notion of periodicity to other types of regular behavior. In this concluding chapter, we briefly touch upon some interesting open questions related to each of the three techniques we explore in this thesis.

In Chapter 2, we described an inherent measurement bias in a common method for measuring the graph-theoretic properties of a network over time, in the presence of even small amounts of missing temporal data. Since most network datasets lack a complete temporal history, this can be a significant problem when the observed (but not necessarily true) trends in graph theoretic properties are used for model building or algorithmic optimizations, for example. We were able to show this bias through Monte Carlo simulations, and conjecture that there is no general way to correct it. This leads to quite a few open questions:

- *Are real dynamic network measurements using the growing network method accurate?* This is probably the most significant open question, and might require either looking for secondary markers of the bias caused by a missing temporal history, domain-specific information, or more advanced models of the dynamics of networks that take both repeated discovery of edges and vertices, as well as the addition of new edges and vertices, into account.

- *Are there network properties that are not affected by the missing data bias?* Almost all the network properties analyzed in the literature exhibit some form of bias with a small amount of missing data. However, spectral properties seem more resilient to these biases. An analytical exploration of which network measures converge to their true trends would be extremely valuable.

- *How else can we measure network properties?* Given that there are serious issues with missing data in real networks, can alternative sampling schemes be developed to give a more accurate picture of the evolution of network properties? For example, to measure trends in the average shortest path length in a network over time, would a sampling strategy that only considers shortest paths between specific pairs of nodes yield a more accurate picture of network evolution than measuring the entire network?

In Chapter 3, we proposed a method for finding both the important periodicities as well as periodic patterns in a dynamic network. In particular, our problem formulation is the first one that aims to describe periodic patterns without redundancy. As a result, we proved that the enumeration and counting variants of our mining problem are in the complexity class P, unlike several other periodic pattern mining formulations that are either #P-complete for counting, or NP-hard for enumeration. We described an efficient algorithm to mine these patterns and periodicities, and found a number of very intuitive periodicities in real datasets: e-mail traffic has principal periodicities of 1 day and 7 days, web server access logs have their highest peaks at 1 day, 7 days, and then 2 and 3 days in decreasing order. Unsurprisingly, repeated patterns of celebrity sightings occur most frequently approximately every 365 days.

Given the success of our algorithm at detecting periodicities in a number of diverse datasets, there are a number of algorithmic and theoretical extensions that can be made:

- *How can approximate periodicity be quantified?* We proposed a heuristic mechanism to analyze approximate periodicities, but did not analyze it theoretically. A strict combinatorial definition of approximate periodicity is likely to be difficult, so statistical approaches might present a better alternative.

- *Can we develop optimal, approximate, parallel or streaming versions of the mining algorithm?* These are natural extensions of the mining algorithm we proposed.

- *Can we detect communities by finding periodic semi-patterns?* A semi-pattern is one which does not appear in its entirety at each timestep. For example, a group of individuals might meet on a strictly periodic schedule, but not all members of the group will attend every meeting. Can these 'visitors' be detected by association with the members? One way to achieve this would be to augment networks with pseudo-nodes for periodic patterns, with edges to all vertices that are part of that pattern, and then apply static graph link prediction to infer which other nodes are likely to be a part of the pattern.

Finally, in Chapter 4, we described the mining of temporally coupled edges, where the occurrence of an edge at a particular time predicts the occurrence of another edge at a later time. We used edge dependencies defined by the line graph transformation of a graph, the first-order differenced time series of edge occurrence times, and a novel Hidden Markov Model formulation for this task. We were able to show that a number of edges in real networks are predictable to a high degree, and that the network structure of these edges is non-trivial, suggesting the existence of a 'predictable core' sub-network. The most prominent open question relates to increasing the accuracy of the predictive models used.

- *Can predictive models for coupled edges be made more accurate?* The HMM model we describe is a step in this direction, and extensions of HMMs such as Hidden Semi-Markov Models [Salfner and Malek, 2007], or even completely different predictive models, could boost the number of coupled edges that can be found.

- *What are other applications for coupled edges?* In Chapter 4, we demonstrated a number of practical applications for mining coupled edges. It would be interesting to see the extent to which they can be found in other datasets, and what temporally predictive edges correspond to.

- *What is the network structure of coupled edges?* The network structure of these edges, in aggregate, would also be illuminating should it have a non-trivial structure or span a large portion of the vertices in the network. We conjecture, based on the limited experiments with our datasets, that such a backbone of predictable interactions exists in information networks; what does this sub-network correspond to in reality, and how effectively does it connect the network?

There are many other interesting open questions, and the increasing quantity and diversity of network data makes robust analytical tools indispensable. This thesis is a small step in the development of a standard suite of such tools.

Bibliography

[Salfner and Malek, 2007] Salfner, F. and Malek, M. (2007). Using hidden semi-markov models for effective online failure prediction. pages 161–174, Los Alamitos, CA, USA. IEEE Computer Society.