Attributed Graph Models: Towards the Sharing of Relational Network Data

Joseph J. Pfeiffer III[1], Sebastian Moreno[1], Timothy La Fond[1],
Jennifer Neville[1] and Brian Gallagher[2]
Purdue University[1] and Lawrence Livermore National Laboratory[2]
{jpfeiffer,smorenoa,tlafond,neville}@purdue.edu, bgallagher@llnl.gov

The growth of the internet has created large scale collections of *relational* data. In these cases, datasets contain relationships between the items or individuals that are being modeled – e.g, hyperlinks connect webpages on the internet, while friendships (Facebook), followers (Twitter) or messages (Email) form between individuals in social networks. Individuals connected through these relationships exhibit *relational correlation*, or a statistical dependence of their attributes [2]. Modeling these relationships can give better predictions about users, or a better understanding of the underlying social processes.

The field of *Statistical Relational Learning* (SRL) utilizes these relational connections to collectively predict the unknown labels in the network, with resulting methods able to largely outperform traditional independent learning methods (for a summary of SRL see [3]). The developed methods can undertake tasks such as identifying fraudulent securities traders or inferring gene interactions, as well as predict user traits or personalize content. Similarly, *Social Network Mining* (SNM) mines information of the individuals given their attributes and relational structure, focusing on tasks such as predicting future links or on identifying communities within a network (for an introduction see [1]). In these domains, large scale data is necessary to drive further research towards developing accurate and scalable algorithms that continuously push the state of the art forwards.

However, attributed data in relational domains is particularly sensitive in comparison to other domains. Datasets such as the UCI collection[1] exist for moderate testing and comparison of traditional machine learning algorithms, while large scale unattributed network repositories exist such as SNAP[2] or the UF Sparse Matrix Collection[3]. In contrast, attributed relational datasets are typically the product of collections of social interactions, such as Facebook, Twitter, LinkedIn, LivingSocial, Email, and more. A large collection of labels for websites exists through the Open Directory Project[4], but requires crawling millions of pages. As the pages are under copyright by their original publishers, these crawls cannot be released. Thus, attributed networks are closely guarded for both copyright, proprietary and privacy reasons. As a result, although public attributed datasets do exist (e.g., IMDB[5] or SNAP's Amazon copurchases), they are rare, small and/or can not be easily distributed.

For many tasks, the exact proprietary information is not needed by the researchers; rather, networks with similar network structure and attribute correlations that capture salient characteristics of the networks would suffice. Advances in understanding and learning on the similar datasets can then translate to successes on the private data. For example, a network with similar (e.g.) clustering, degree distributions and attribute correlations with a billion vertices could allow for demonstrations of algorithm scalability, with the resulting methods translated and implemented on datasets such as Facebook or Twitter. In this case, the actual Facebook or Twitter network is not needed, simply a reasonable substitute in terms of size and structure.

Recent advances in generative network models [6, 7] allow for scalable *learning* and *sampling* graph structure. By making reasonable restrictions on the search space, these methods can sample from the space of edges in subquadratic time (in terms of the number of vertices). This allows them to scale to networks with billions of vertices and accurately capture the underlying network structure. However, the assumptions they make are carefully crafted to allow for scalable learning and sampling of real world network structure, which can not incorporate vertex attribute information.

Our recently proposed *Attributed Graph Models* (AGM) is the first step to solving this problem [8]. AGM extends any existing *scalable* generative graph models to incorporate attributes on the vertices. In doing so, AGM provably samples from the joint distribution of attributes and edges, with both the sample and model then available for other researchers to use. AGM maintains the structural characteristics provided by the graph models and incorporates the attribute dependencies, while remaining subquadratic in runtime.

In particular, AGM generalizes and exploits a common key structural assumption of generative network models. Namely, generative network models incrementally sample an observed edge from all possible edges and insert it into the generated network. This process repeats until enough edges are inserted into the network. AGM augments this pro-

[1] archive.ics.uci.edu/ml/
[2] snap.stanford.edu
[3] www.cise.ufl.edu/research/sparse/matrices/index.html

[4] www.dmoz.org
[5] www.imdb.com

KDD at Bloomberg 2014 New York, NY

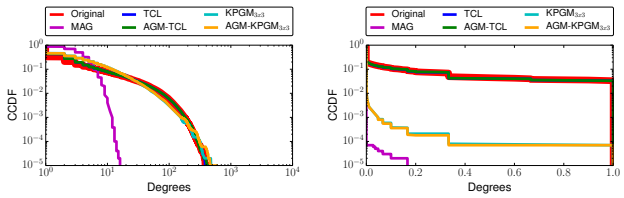

Figure 1: Degree distributions and Clustering Coefficient distributions for TCL, AGM-TCL, KPGM and AGM-KPGM. The AGM models capture the generative graph model's structure for both.

Model	Facebook Correlations		
	R	P	RP
Original	0.108	.211	0.106
MAG	0.584	0.436	0.002
TCL	0.001	0.001	0.001
AGM-TCL	**0.128**	**0.219**	**0.093**
$KPGM_{3x3}$	0.001	-0.001	0.001
$AGM\text{-}KPGM_{3x3}$	**0.132**	**0.221**	**0.092**

Table 1: Correlations for attributes in each dataset. Bold indicates within .05 of the original network.

cess by selectively inserting some edges, and omitting others. The edges chosen for insertion reflect the conditional distribution of the edges given the attributes, modeling the dependencies between the edges and attributes.

The AGM process begins by learning independent distributions for both the attributes (using statistical models [5]) and edges (using a generative network model), then, samples a new set of attributes for the vertices. AGM uses the learned generative graph model to sample a network independent of the attributes, and measures the correlations observed in comparison to the observed attribute edge dependencies in the true network. AGM then uses the differences between these correlations to construct acceptance probabilities. Lastly, AGM discards the network sample and constructs a new network by repeatedly sampling from the generative graph model, but selectively rejecting proposed edges based on the endpoint attributes using the learned acceptance probabilities.

As a demonstration, we apply AGM on a Facebook network with 444,817 vertices and 1,016,621 edges. We test AGM combined with two generative graph models: the first is KPGM with a 3x3 initiator matrix, the second is the *Transitive* Chung Lu (TCL) model (a variant of CL). We choose to have AGM model the joint distribution of two attributes: Political Views and Religious Views. The structural characteristics (degree and clustering) of the generated networks are given in Figure 1: we see that AGM-TCL is structurally identical to TCL, while AGM-KPGM is structurally identical to KPGM. However, Table 1 shows the correlations produced by the AGM models for the Politics and Religion features match the correlations of the original data, while the baseline generative network models (KPGM and TCL) do not match the correlation of the original network. AGM models not only the individual correlation of Religion and Politics, but also the cross correlation (the RP column). We augment the Multiplicative Attributed Graph (MAG) [4] model for comparison; however, this model is intended for inferring latent attribute characteristics and cannot effectively model the observed correlations.

The abilities of AGM are further highlighted in the example. First, approximately 6,000 people have labels within the Facebook network, yet AGM is able to sample a network with nearly 500,000 vertices and over 1,000,000 edges. This larger attributed network allows for more extensive testing of relational algorithms on scalable datasets. Second, the randomly generated network contains the same structural and attribute characteristics as the original Facebook network, but is distinct. To date, we have used AGM to release synthetic representations of five original attributed networks. We are actively in the process of releasing AGM code, so others can utilize it to release synthetic networks with similar characteristics as proprietary networks[6].

The released AGM datasets represent a new direction towards testing and reproducibility efforts in SRL and SNM research. However, there is still considerable work to undertake. The current AGM method uses simple discrete multinomials to represent the acceptance probabilities. Future efforts should focus on more advanced modeling of the distribution of attributes given edges, with considerable focus on accurate statistical models of the edge relationship structure. Further, theoretically the accept-reject processes hold for continuous variables, but we have not yet investigated this domain. Lastly, high dimensional attribute vectors could impact the acceptance rates of the AGM process. Future efforts should focus on overcoming this possible limitation through methods such as annealing. By creating generative models of attributed network data, we avoid the limitations of proprietary relational data to advance SRL and SNM and allow for further study in large scale domains.

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This research is supported by NSF under contract numbers IIS-1149789, CCF-0939370 and IIS-1219015.

References

[1] C. C. Aggarwal, editor. *Social Network Data Analytics*. Springer, 2011.

[2] L. Fond and Neville. Randomization tests for distinguishing social influence and homophily effects. WWW '10.

[3] L. Getoor and B. Taskar. *Introduction to Statistical Relational Learning*. The MIT Press, 2007.

[4] M. Kim and J. Leskovec. Multiplicative attribute graph model of real-world networks. *Internet Mathematics*, 8(1-2):113–160.

[5] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[6] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *J. Mach. Learn. Res.*, 11:985–1042, Mar. 2010.

[7] J. J. Pfeiffer III, T. La Fond, S. Moreno, and J. Neville. Fast generation of large scale social networks while incorporating transitive closures. In *SocialCom*, 2012.

[8] J. J. Pfeiffer III, S. Moreno, T. La Fond, J. Neville, and B. Gallagher. Attributed graph models: Modeling network structure with correlated attributes. In *WWW*, 2014.

[6] nld.cs.purdue.edu/agm