Attributed Graph Models: 
Towards the Sharing of Relational Network Data

Joseph J. Pfeiffer III\textsuperscript{1}, Sebastian Moreno\textsuperscript{1}, Timothy La Fond\textsuperscript{1}, Jennifer Neville\textsuperscript{1} and Brian Gallagher\textsuperscript{2}

Purdue University\textsuperscript{1} and Lawrence Livermore National Laboratory\textsuperscript{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 \cite{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 \cite{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 \cite{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\textsuperscript{1} exist for moderate testing and comparison of traditional machine learning algorithms, while large scale unattributed network repositories exist such as SNAP\textsuperscript{2} or the UF Sparse Matrix Collection\textsuperscript{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\textsuperscript{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 \cite{5, 6}, 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 \cite{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 \cite{8}. 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 \cite{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-

\textsuperscript{1}archive.ics.uci.edu/ml/

\textsuperscript{2}snap.stanford.edu

\textsuperscript{3}www.cise.ufl.edu/research/sparse/matrices/index.html

\textsuperscript{4}www.dmoz.org

\textsuperscript{5}www.imdb.com

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD at Bloomberg 2014 New York, NY
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.
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.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Facebook Correlations</th>
<th>R</th>
<th>P</th>
<th>RP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td></td>
<td>0.108</td>
<td>-0.211</td>
<td>0.106</td>
</tr>
<tr>
<td>MAG</td>
<td></td>
<td>0.584</td>
<td>0.436</td>
<td>0.002</td>
</tr>
<tr>
<td>TCL</td>
<td></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>AGM-TCL</td>
<td></td>
<td>0.128</td>
<td>0.219</td>
<td>0.093</td>
</tr>
<tr>
<td>KPGM$_{3\times3}$</td>
<td></td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>AGM-KPGM$_{3\times3}$</td>
<td></td>
<td>0.132</td>
<td>0.221</td>
<td>0.092</td>
</tr>
</tbody>
</table>

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


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 networks⁹.

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 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


⁹nld.cs.purdue.edu/agm