Chapter 10
Extracting and Measuring Relationship Strength in Social Networks

Steven Gustafson
GE Global Research, USA

Abha Moitra
GE Global Research, USA

ABSTRACT

This study examines how extracting relationships from data can lead to very different social networks. The chapter uses online message board data to define a relationship between two authors. After applying a threshold on the number of communications between members, the authors further constrain relationships to be supported by each member in the relationship also having a relationship to the same third member: the triangle constraint. By increasing the number of communications required to have a valid relationships between members, they see very different social networks being constructed. Authors find that the subtle design choices that are made when extracting relationships can lead to different networks, and that the variation itself could be useful for classifying and segmenting nodes in the network. For example, if a node is 'central' across different approaches to extracting relationships, one could assume with more confidence that the node is indeed 'central'. Lastly, the chapter studies how future communication occurs between members and their ego-networks from prior data. By increasing the communication requirements to extract valid relationships, it is seen how future communication prediction is impacted and how social network design choices could be better informed by understanding these variations.

DOI: 10.4018/978-1-61350-444-4.ch010

INTRODUCTION

Studies on real-world complex social networks have not typically considered the existence of erroneous links in the observed social network. Most prior work assumes that with a large enough sample size, the true, or accurate, relationships will out-weigh the false ones. However, as demonstrated in two recent publications (Gustafson et al., 2009, Choudhury et al., 2010a), this is not always the case for real-world social network data. As more data becomes available, and more societal and cultural benefits of analyzing social networks are identified (Lazer et al., 2009), it will be critical to fully understand social network data issues. In this paper, we build upon recent work (Gustafson et al., 2009, Choudhury et al., 2010a) that shows how a simple threshold condition on the number of ties between nodes can lead to dramatic effects in the network structure as well as the types of information one could infer about the network and individual nodes in the network.

BACKGROUND

Recently, Latapy and Magnien (2008) validate the sample size assumption and show that it is possible to distinguish between cases where this assumption is reasonable, those large enough data sets overcome issues of noise within the data, and they also find cases where the assumption must be discarded. Latapy and Magnien (2008) conclude that the qualitative properties of some statistics do not depend on the sample size, as long as it is not trivially small. They find that some statistics, like average degree, can be used to infer other statistics, whereas other statistics like transitivity are generally unstable as sample sizes grow. These more 'structural' statistics are somehow more related to other measures like maximal degree. While qualitative estimations of the more stable statistics, for example average degree, are possible, obtaining accurate estimations of these statistics remains difficult. Lin and Zhao (2005) present a study on the impact of erroneous links on degree distribution estimation and show that the degree distributions of power-law networks still have power-law degree distributions for the middle range degrees, but can be greatly distorted for low and high degrees. Borgatti et al. (2006) show that centrality measures are surprisingly similar with respect to pattern and level of robustness to data errors and different types of errors have relatively similar effects on centrality robustness. The limitation of this last study is that they consider only random errors on random networks. As we are primarily interested in real-world data and the impact of design choices when extracting social network relationships, we will not address prior work that has examined this topic within simulated data. In Costenbader and Valente (2003), several centrality measures are studied for stability across several different social network data sets. Using various degrees of sampling, the authors find that some measures are more stable than others, and some, like Bonacich's Eigenvector measure, are very unstable when comparing the correlation between the centrality measurements of the sampled populations to the centrality measurement of the original un-sampled population. Recently, in Choudhury et al. (2010b), sampling is studied on very large-scale data to estimate information diffusion. The authors find that sampling can be improved by using contextual information, like physical location, to direct future samples toward other actors that may share similar interests or attributes and be important for the sample to estimate information diffusion.

While some statistics and social network measurements are more stable across sampling or collection strategies, we still face a lot of uncertainty when considering the future, where data is collected for other purposes and then leveraged in a social network analysis setting. Marsden (1990) discusses several issues with social network data collection that are very relevant today, including: the fact that some relationships are persistent