Who Matters? Finding Network Ties Using Earthquakes

Siddhartha Basu, David Daniels, Anthony Vashevko

December 12, 2014

1 Introduction

While social network data can provide a rich history of personal behavior and interpersonal interactions, it is often unclear how high quality data maps onto "real" social constructs. Deep trace data (e.g. emails, cell phone logs) provides thorough measurement, but cannot often identify important aspects of behavior: [3] is a person emailing their friend, their spouse, their colleague? Because the content of such relationships can be of extreme importance to social theory [4], researchers have looked for ways to identify such ties in the absence of true data, often relying on tie weighting. Using a trace dataset that allows for the identification of close relationships, we attempted to uncover network characteristics that predict close relationships, as well as to quantify the granularity of network data necessary to be able to perform this prediction reliably.

The attempt was not successful. We attempted to predict the presence of two kinds of close relationships using support vector machines (SVMs) and logistic regression in networks. Optimal prediction models failed to predict any important relationships in the test set. That is to say, though models were occasionally able to separate unimportant ties from important ties during training, they predicted that all test ties were unimportant to the focal ego.

These results are naturally disappointing – they suggest that standard network measures of tie importance are incredibly coarse for the kind of social science conducted on trace data. The conclusion of the paper discusses various flaws of the approach we took here, suggesting ways to improve this analysis in a way that might confirm the intutitions behind network measures. Nevertheless, our results appear to be disappointing for researchers who wish to rely on trace data as a high quality measure of social life.

2 Data

This paper relies on a novel dataset of cell phone calls and texts acquired from a Chinese telecommunication company. The data covers phone use in a single province (Sichuan) over a period of four months (March-June 2013). The data includes information on all calls and texts made by or two a population of 157 thousand subscribers, along with basic demographic information about these subscribers, such as phone and plan characteristics.

Two features set this data apart: First, the data includes information about family plans, allowing us to differentiate communications between family members from communications to outside partners. Second, a major earthquake struck the province during the period of observation. Following other ongoing work on this dataset [5], we propose that the first calls and texts made after the earthquake reach out to important individuals – key providers of support and resources (emotional and otherwise) to the subscribers in the data set. These people represent a unique ground truth for identifying the contacts that matter to people. Our goal in this project is to attempt to predict these two classes of important ties: family members, and targets and sources of first calls.

For each ego, we identify the first call or text made to or from that ego, and label the source or target of that communication as the first call partner. One weakness of this approach is that we examine only first calls. It would of course be straightforward to extend this measure to be a ranking of alters by their ordinal 'importance' to the ego.

A more serious shortcoming of this dataset is that the Chinese telecommunications market is split into several major firms. Because our data comes from only one of those firms, we have no knowledge of the communications among alters – we do not have the complete network of communication, and in fact we do not have complete communication networks even for the immediate ego-network around any subscriber. As such, we aren't able to use measures of network embeddedness to identify important ties.

3 Features

We used several classes of data to represent the kind of information that might be available to researchers. We examine tie weight, coarse network timing information, and finer network timing information. We restrict data to all communication that took place prior to the time of the earthquake.

Tie Weight Tie weight is the overall number of times that i contacted j.

Tie Variance Tie variance is the week-to-week variance in the number of communication events that occured from i to j.

Interresponse Time We calculated the median time elapsed between the last communication from j to i and the next event from i to j. Because the average inter-response time is essentially the inverse of the frequency of communication events, we used the median to capture the extent to which communication events clustered between the two agents. As an alternative measure of the same concept, we calculated the fraction of all inter-response times that took place within an hour of each other.

Tie weight and variance were split between voice calls and text, and were additionally split for ego-to-alter communication and alter-to-ego communication. That is to say, we had separate features for the amounts of texts a subscriber sent to some communication partner and for the number of texts the partner sent back. Finally, because several of these measures were undefined in the absence of sufficient communication from i to j, we included indicators for missing data as hopefully informative features.

4 Models

To capture the idea that researchers might have limited information available in any given data set, we nested the above information in three models:

Model 1 Model 1 included tie weight features only, to represent a barebones network dataset.

Model 2 Model 2 included all features from Model 1 as well as tie variance features, to capture the availability of limited network timing data.

Model 3 Model 3 included all features from Model 2 as well as the response time features, to represent a rich network timing dataset.

Our data analysis was conducted in R, using the e1071 wrapper to aid us in implementing libsvm. Libsvm is a powerful package that provides extensive flexibility in classification algorithms and analysis which proved to be useful in applying the following learning models and then studying how they performed. [2]

Logistic Regression Our first cut at the analysis is by implementing logistic regression on the data. Using Rs base package, we ran logistic regressions to predict one of two dependent variables. The first of these is an indicator for whether two individuals are in the same family plan, and the second of these is an indicator for whether there was a call between two people in the immediate aftermath of the earthquake. The independent variables are those described above. Specifically, the first model includes call volume, the second includes volume and call variance, and the third includes the second combined with time between calls and texts, and how long it takes people to respond to calls and texts between egos and alters.

Support Vector Machine We then proceeded to analyze the data using support vector machines. A main advantage of SVMs is that by using kernels, we can gain insights from higher dimensional spaces while still using a relatively small number of features. Our first SVM implemented a Gaussian kernel. We did this by utilizing LIBSVM's radial basis function (RBF). The outcome variables (same family/not, and call after earthquake/not) were the same as in the logistic regression. The same can be said of our features. The second kernel that we used was the sigmoid function. The outcome variables and features are again the same as before.

Finally, our modeling strategy assumes the independence of ties. We select samples for training and testing by selecting a random subset of subscribers and including for analysis all alters who called or were called by these subscribers. There is a problem here in that one of our measures of importance, the first call, is by definition constrained to a single alter for any given ego – any person had to have exactly one first call. We did not incorporate this constraint into our model. Doing so would appear to lead towards a relatively complicated approach of multinomial classification with a variable number of 'classes' per ego. Given more time, this approach would be worth exploring. Nevertheless, our hope was that the current approach would at best produce a high rate of false positives, as the model identified important alters that were not first calls, but came close.

5 Results

5.1 SVM Parameter Selection

We attempted to select optimal parameter settings for two classes of SVM via a grid search procedure. We fit models to a subset of the data and tested them on another subset. Our optimality criteria were models with minimal RMSE and maximal recall. We selected the RMSE criterion in order to capture the overall accuracy of the model. The number of positive examples was very small, however – egos had about 100 possible contacts on average, and only about 2-3 of these were family ties, and only one of these was a first call tie. Because we were interested in identifying instances of these rare important ties, we also chose to maximize recall.

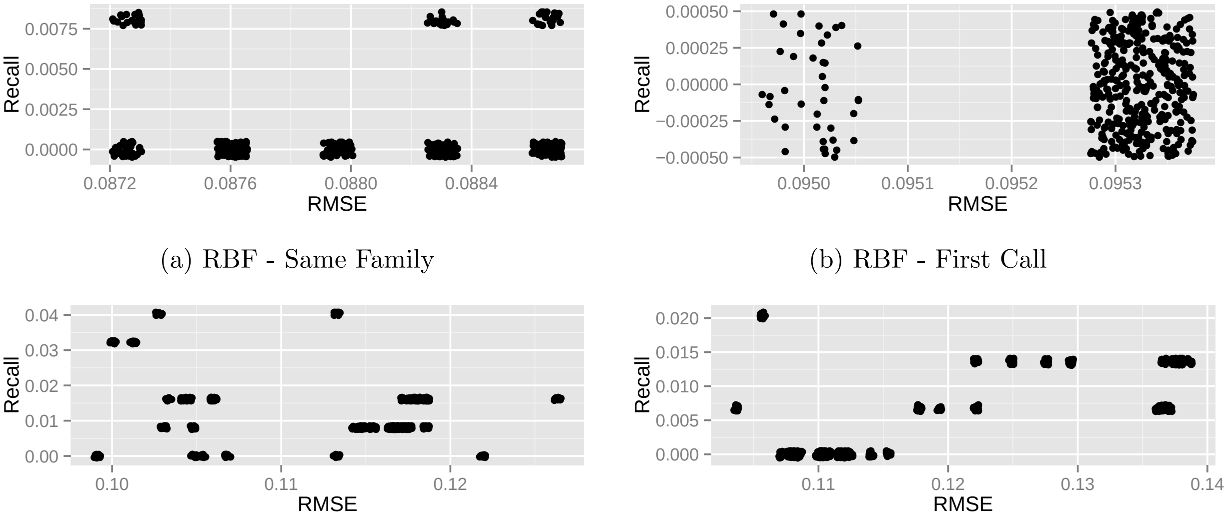
Figure 1 shows the outcome of these tests. The plots are jittered because almost all parameter values produced identical prediction outcomes: the major difference was how many positive predictions a given model was able to make. There are two points to note here: First, RBF outperformed the sigmoid kernel on RMSE measures, but the sigmoid kernel showed better recall and greater variability of recall across both family and first call models. Second, RBF was completely unable to predict instances of first callers.

With these results in mind, we proceeded to the full models by selecting the grid parameters that produced the lowest RMSE at the highest observed recall – essentially, those models that had the fewest false positives of those that had the most true positives. We tried out different combinations of these 'optimal' results, but this did not substantially affect results from the final models.

5.2 Full Models

As explained above, we estimated the three models above with three techniques: logistic regression, SVM with a Gaussian kernel, and SVM with a sigmoid kernel. Because solution time for the SVM algorithm grows quickly in the size of the data set, we restricted the training and test sets for all models to a 0.2 percent sample of all subscribers. This about 30,000 tie observations for about 300 subscribers. Table 1 presents the RMSE and recall on the test set after fitting all models.

As the table suggests, most models were unable to predict any true positives. Table 2 presents a sample confusion table for the logistic



(a) RBF - Same Family

(b) RBF - First Call

(c) Sigmoid - Same Family

(d) Sigmoid - First Call

Figure 1: SVM Parameter Tuning: Recall and RMSE

model on same family, which was the most successful model for producing positive predictions. As the confusion table suggests, the RMSE and overall accuracy were substantially driven by the true incidence rate of same family or first call ties – because the models classified almost all ties as non-instances of the same family or first call class, they misclassified all of these as false negatives. In fact, all SVM models made no positive predictions, failing completely to identify important ties in the test set.

6 Discussion

There are several concerns with the models presented here. First, it may well be that our features imperfectly captured the true nature of the network. This is troubling – at the very least Model 1 measuring total tie communications should have helped discriminate between important and unimportant ties. In fact, it did show significant predictive power in the logistic regression. Nevertheless, it did not help discriminate among observations in the test set.

It may well be that our modeling strategy failed to capture ego-level heterogeneity in communication: people with radically different volumes of communication may have confused the models. It may have been better to measure tie strengths relative to the ego's average or total volumes of communication.

The more troubling possibility is that dependence among ties made binary classification unviable – no model that doesn't take into consideration the overall network of ties around an ego could successfully make same family or first call predictions. Given our inability to observe most members of the network, however, it is unclear how far we could have gone down this route.

7 Conclusion and Future

The most surprising findings are that network information does not appear to easily identify important ties, and perhaps more importantly, that adding deeper information appears to have little if any discriminatory benefit in this task. This is troubling for existing studies of social networks, as well as for the research promise of trace data.

The broad conclusion appears to be that important ties hide well in networks – it is not clear that detailed network data is capable of identifying family members or important personal and support ties. SVMs in particular seem to have trouble making reasonable predictions. As such, it remains unclear the extent to which standard

Same Family		
	RMSE	Recall
Logistic (M1)	0.0874	0.0450
Logistic (M2)	0.0876	0.0450
Logistic (M3)	0.0876	0.0450
SVM-RBF (M1)	0.0859	0
SVM-RBF (M2)	0.0859	0
SVM-RBF (M3)	0.0859	0
SVM-Sigmoid (M1)	0.0859	0
SVM-Sigmoid (M2)	0.0859	0
SVM-Sigmoid (M3)	0.0859	0
First Call		
	RMSE	Recall
Logistic (M1)	0.0988	0
Logistic (M2)	0.0990	0
Logistic (M3)	0.0990	0
SVM-RBF (M1)	0.0988	0
SVM-RBF (M2)	0.0988	0
SVM-RBF (M3)	0.0988	0
SVM-Sigmoid (M1)	0.0988	0
SVM-Sigmoid (M2)	0.0988	0
SVM-Sigmoid (M3)	0.0988	0

Table 1: RMSE and Recall of Full Models

network measures of tie importance are effective.

Further research could further improve on our methods, and might include the following tasks:

1. Implement SVMs on a larger subsample of the dataset

2. Explore alternate estimation processes, perhaps moving to Hadoop MapReduce to allow us to process both the volume of data and an extended set of features.

3. Extend analysis to other exogenous shocks to cell phone communications, such as other natural disasters, [1] surprise financial news, [7] unanticipated lapses in cell phone service, etc.

4. Extend to analysis of exogenous shocks to other social networks such as Twitter, Facebook, or LinkedIn. [6]

True	0	1
Predicted 0	29868	212
1	19	10

Table 2: Confusion Table: Logistic (M3) for Same Family

References

[1] Joshua Blumenstock, Nathan Eagle, and Marcel Fafchamps. Risk and Reciprocity Over the Mobile Phone Network: Evidence from Rwanda. 2011.

[2] Chih-chung Chang and Chih-jen Lin. LIbsVm: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1—-27:27, 2013.

[3] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36):15274–8, sep 2009.

[4] Mark S. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[5] Jayson J. Jia and Jianmin Jia. Tie Importance and Social Network Embeddedness Revealed by Earthquake. 2014.

[6] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford Large Network Dataset Collection, 2014.

[7] A. Craig MacKinlay. Event Studies in Economics and Finance. *Journal of Economic Literature*, 35(1):13–39, 1997.