

# Modeling Large-Scale Networks Based on Mobility Data

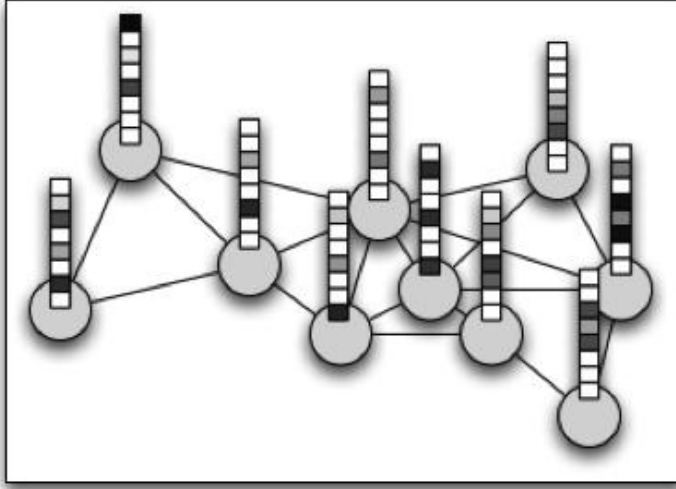
*Tony Jebara*

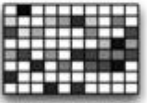
*Columbia University*

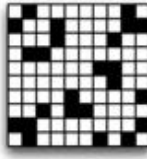
Many real-world social networks contain both topological connectivity information as well as attributes (or features) associated with each node. We consider inferring such networks from social data, mobile data and location data from large populations of users. While most network growth models are based on incremental link analysis [Adamic and Adar, 2003], we explore how users' data profiles alone (without any connectivity information) can be used to infer their connectivity with others. For example, in a class of incoming freshmen students with no known friendship connections, can we predict which pairs will become friends at the end of the year using only their demographic profile information? Similarly, can we use co-location to predict communication? For instance, by observing only the location history from a population of mobile phone users, can we predict what pairs of users are likely to communicate and text/call each other?

To learn how to reconstruct these networks, we present structure-preserving metric learning (SPML) [Shaw, Huang and Jebara, 2011] and degree-distributional metric learning (DDML) [Huang, Shaw and Jebara, 2011]. These are algorithms for learning a Mahalanobis distance metric from a network and profile information. The goal is to learn distance metrics that capture the underlying inherent connectivity structure of the network. SPML learns a metric which is

structure-preserving [Shaw and Jebara, 2009], meaning a connectivity algorithm (such as k-nearest neighbors or b-matching [Huang and Jebara, 2007]) will yield the correct connectivity when applied using the distances from the learned metric.





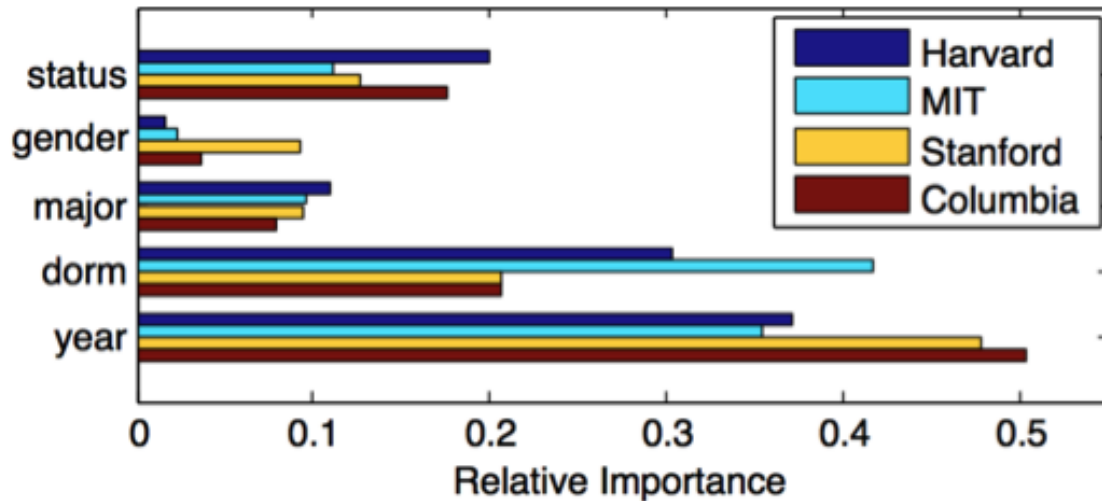
$$\mathbf{X} \in \mathbb{R}^{d \times n}$$


$$\mathbf{A} \in \mathbb{B}^{n \times n}$$

The SPML approach begins with a known network with known attributes for the nodes. For example, we observe a friendship network of students as they graduate from university after their undergraduate program is completed. Assume that this network is represented by an adjacency matrix  $\mathbf{A}$ . We also observe, for each individual in the network, their static demographic attributes (their age, height, weight, home-town, income bracket, favorite music, dorm room assignment, etc.). Let the demographic attributes of all the users be represented by a matrix  $\mathbf{X}$ . We use  $\mathbf{A}$  and  $\mathbf{X}$  as training data to learn our distance metric. After training, our test goal is to predict the adjacency matrix for a new set of incoming students on their first day at university by only observing their demographic attributes  $\mathbf{X}'$ . Our prediction should closely match the true  $\mathbf{A}'$  adjacency matrix which we eventually obtain at graduation time. More specifically,  $\mathbf{A}$  and  $\mathbf{X}$  allow us to learn an appropriate distance

metric to use when computing the distance between user  $i$ 's demographic vector  $\mathbf{x}_i$  and user  $j$ 's demographic vector  $\mathbf{x}_j$ . For example, how much does an age difference between two users matter relative to a height difference when computing the similarity or distance between a pair of users? Once we have found a good metric that balances all the multivariate demographic dimensions, we can reconstruct  $\mathbf{A}$  from  $\mathbf{X}$  in our training data. To then test how well this method performs, we try reconstructing an unseen network  $\mathbf{A}'$  from only new  $\mathbf{X}'$  demographic data.

We evaluate performance using area-under-the receiver-operator-characteristic curve. We show a variety of synthetic and real-world experiments where SPML predicts link patterns from node features more accurately than standard techniques [Shaw, Huang and Jebara, 2011] [Huang, Shaw and Jebara, 2011]. In particular, our approach outperforms simple naive distance metrics (like Euclidean distance), relational topic models [Chang and Blei, 2010] and support vector machine classifiers [Boser, Guyon and Vapnik, 1992]. We further demonstrate a method for optimizing SPML based on stochastic gradient descent which removes the running-time dependency on the size of the network and allows the method to easily scale to networks with hundreds of thousands of nodes and millions of edges. We show how to build such networks from FaceBook data, Wikipedia data, FourSquare data and mobile phone call detail records. Once a network is built, we can do a variety of interesting things with it. These include visualizing the network [Shaw and Jebara, 2009] and predicting unknown labels about the users (marital status, income, and so on) [Wang, Jebara and Chang, 2013]. We describe the graph algorithms to accomplish these tasks as well.



Some interesting interpretable findings emerge. For instance, through FaceBook social network data [Traud, Mucha and Porter, 2011], we find that Harvard students are relatively more picky about differences in relationship status when forming their friendships. Stanford students and Columbia students are relatively more sensitive to differences in graduation year when forming friendships. MIT students are most sensitive to differences in dorm assignments when forming friendships. Thus, social network structure helps us tease apart the space of demographic attributes and determine which demographic differences are more or less relevant.

Just like FaceBook data, mobile phone data is also well-suited to our approach. We consider a large data-set of location-augmented call detail records (CDRs) from a mobile phone carrier. In this data-set, we use phone calls and text messages between pairs of users to establish the existence of a friendship relationship between them (i.e. an edge in the social network graph). The attributes of each user are his or her location history (the places they visited as obtained via GPS or tower-triangulation of their mobile device). Our broad goal is to learn

metrics that show which co-locations matter more for predicting friendship? In other words, which meeting places are more likely to correlate with (or predict) communication. Are people more likely to be friends if they spend time together in high population density regions or low population density regions? Are people more likely to be friends if they co-locate in a coffee shop or if they co-locate at a subway station?

In conclusion, the graph topology of a social network helps redefine our metrics of similarity and dissimilarity and reshapes the axes of demographic dimensions. With the appropriate algorithms, it can elucidates what specific aspects of user profiles and demographics are predictive of communication and social interaction [Newman 2003] [Lazer et al. 2009].

## References

J. Wang, T. Jebara and S.F. Chang. "Semi-Supervised Learning Using Greedy Max-Cut". Journal of Machine Learning Research (JMLR), 14(Mar):771-800, 2013.

B. Shaw, B. Huang and T. Jebara. "Learning a Distance Metric from a Network". Neural Information Processing Systems (NIPS), December 2011.

B. Huang, B. Shaw and T. Jebara. "Learning a Degree-Augmented Distance Metric From a Network". Beyond Mahalanobis: Supervised Large-Scale Learning of Similarity Workshop, Neural Information Processing Systems (NIPS), December 2011.

B. Shaw and T. Jebara. "Structure Preserving Embedding". International Conference on Machine Learning (ICML), June 2009.

T. Jebara, J. Wang and S.F. Chang. "Graph Construction and b-Matching for Semi-Supervised Learning". International Conference on Machine Learning (ICML), June 2009.

B. Huang and T. Jebara. "Exact Graph Structure Estimation with Degree Priors". International Conference on Machine Learning and Applications (ICMLA), December 2009.

D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. "Computational Social Science". Science, February 6 2009.

B. Huang and T. Jebara. "Loopy Belief Propagation for Bipartite Maximum Weight b-Matching". Artificial Intelligence and Statistics (AISTATS), March 2007.

J. Chang and D. Blei. "Hierarchical Relational Models for Document Networks". Annals of Applied Statistics, 4:124-150, 2010.

B.E. Boser, I.M. Guyon and V.N. Vapnik. "A Training Algorithm for Optimal Margin Classifiers". Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, 1992.

A. Traud, P. Mucha, and M. Porter. Social Structure of FaceBook Networks. CoRR, abs/1102.2166, 2011.

M. Newman. The Structure and Function of Complex Networks. SIAM REVIEW, 45:167-256, 2003.

L.A. Adamic and E. Adar. Friends and Neighbors on the Web. Social networks, 2003.