An Evolutionary Algorithm Approach to Link Prediction in Dynamic Social Networks

Catherine A. Bliss, Morgan R. Frank, Christopher M. Danforth, & Peter Sheridan Dodds
Computational Story Lab, Department of Mathematics & Statistics, Vermont Complex Systems Center, & Vermont Advanced Computing Core

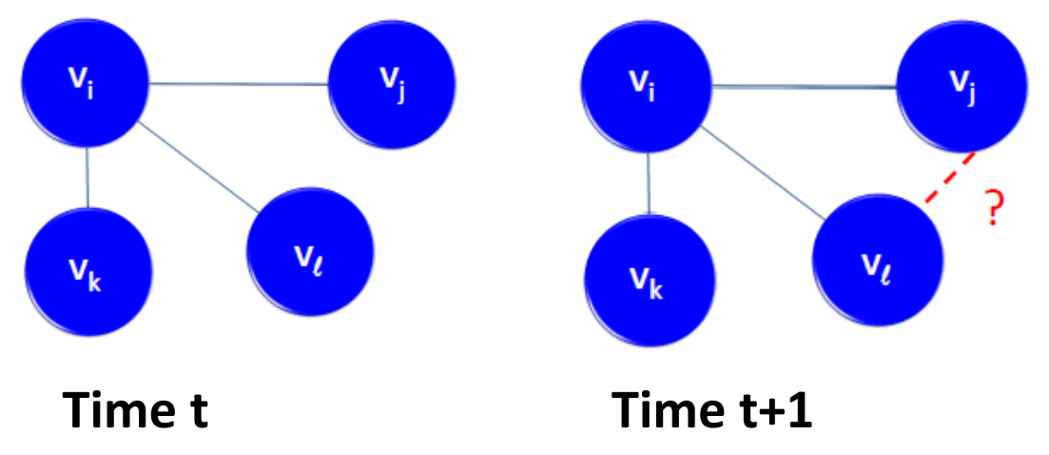
Abstract

Many real world, complex phenomena have an underlying structure of evolving networks where nodes and links are added and removed over time. A central scientific challenge is the description and explanation of network dynamics, with a key test being the prediction of short and long term changes. For the problem of short-term link prediction, existing methods attempt to determine neighborhood metrics that correlate with the appearance of a link in the next observation period. Recent work has suggested that the incorporation of user-specific metadata and usage patterns can improve link prediction, however methodologies for doing so in a systematic way are largely unexplored in the literature. Here, we provide a novel approach to predicting future links by applying an evolutionary algorithm (Covariance Matrix Adaptation Evolution Strategy) to weights which are used in a linear combination of sixteen neighborhood and node similarity indices. We examine Twitter reciprocal reply networks constructed at the time scale of weeks, both as a test of our general method and as a problem of scientific interest in itself. Our evolved predictors exhibit a thousand-fold improvement over random link prediction and high levels of precision for the top twenty predicted links, to our knowledge strongly outperforming all extant methods. Based on our findings, we suggest possible factors which may be driving the evolution of Twitter reciprocal reply networks.

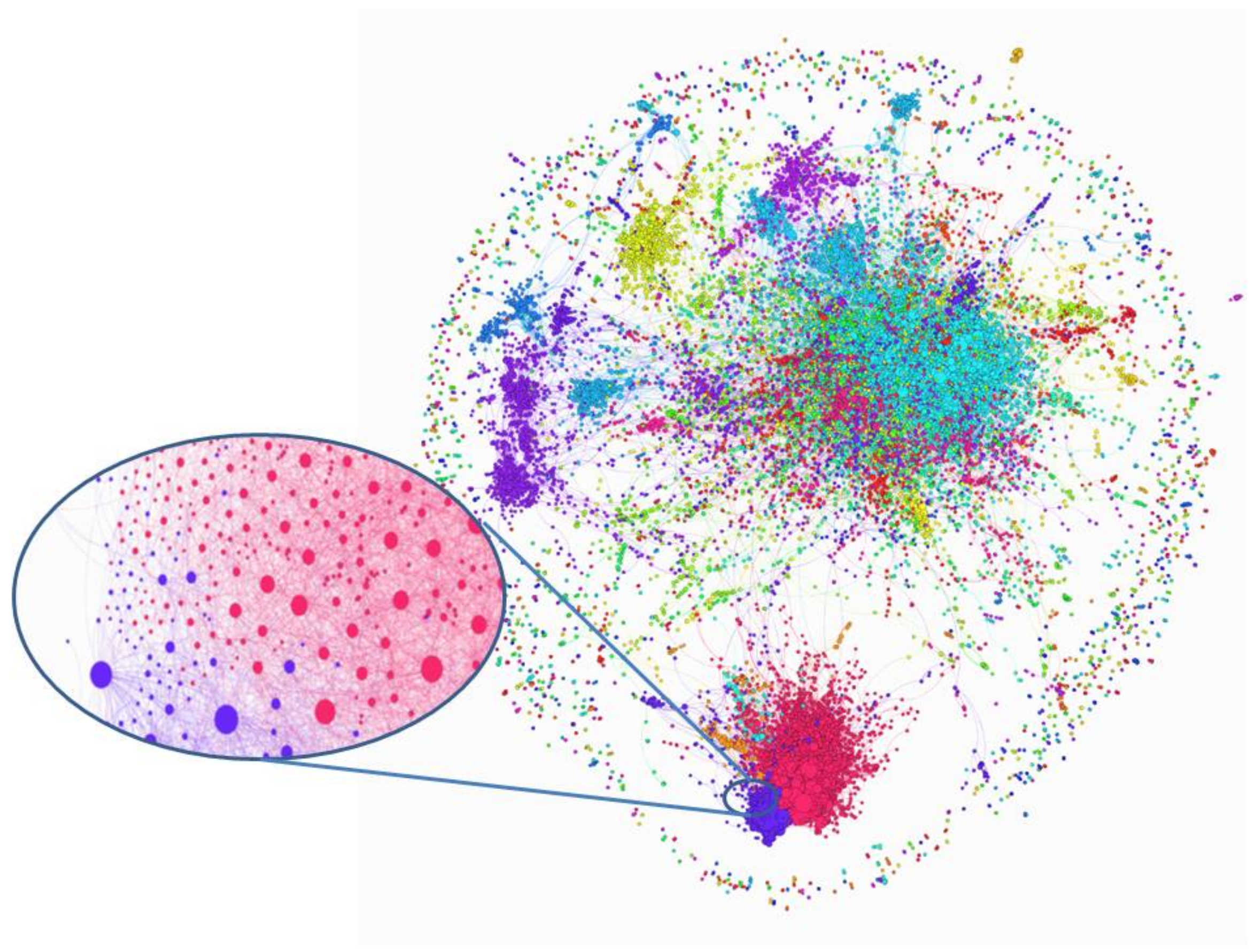
The link prediction problem

Individuals, represented by nodes, may enter or exit the network, while interactions, represented by links, may strengthen or weaken. While network growth models capture global properties, the link prediction problem explores localized dynamics, such as who will be connected to whom in the future.

The link prediction problem: *Given a snapshot of the network at time=t, can we predict links which will appear in time=t+1?*



Reciprocal reply networks



Our data set consists of over 51 million tweets collected via the Twitter API service from 9/9/08-12/1/08. From tweets, we construct reciprocal reply networks (Bliss et al., 2012). The visualization above depicts a Twitter reciprocal reply network constructed from messages sent within 1 week. Nodes represents individuals and links represent evidence of reciprocated replies during the time unit of analysis.

Prediction Outcome		P	N
	P'	True Positives (TP) *hits*	False Positives (FP) *false alarms*
	N'	False Negatives (FN) *misses*	True Negatives (TN) *correctly rejected*
		True State	

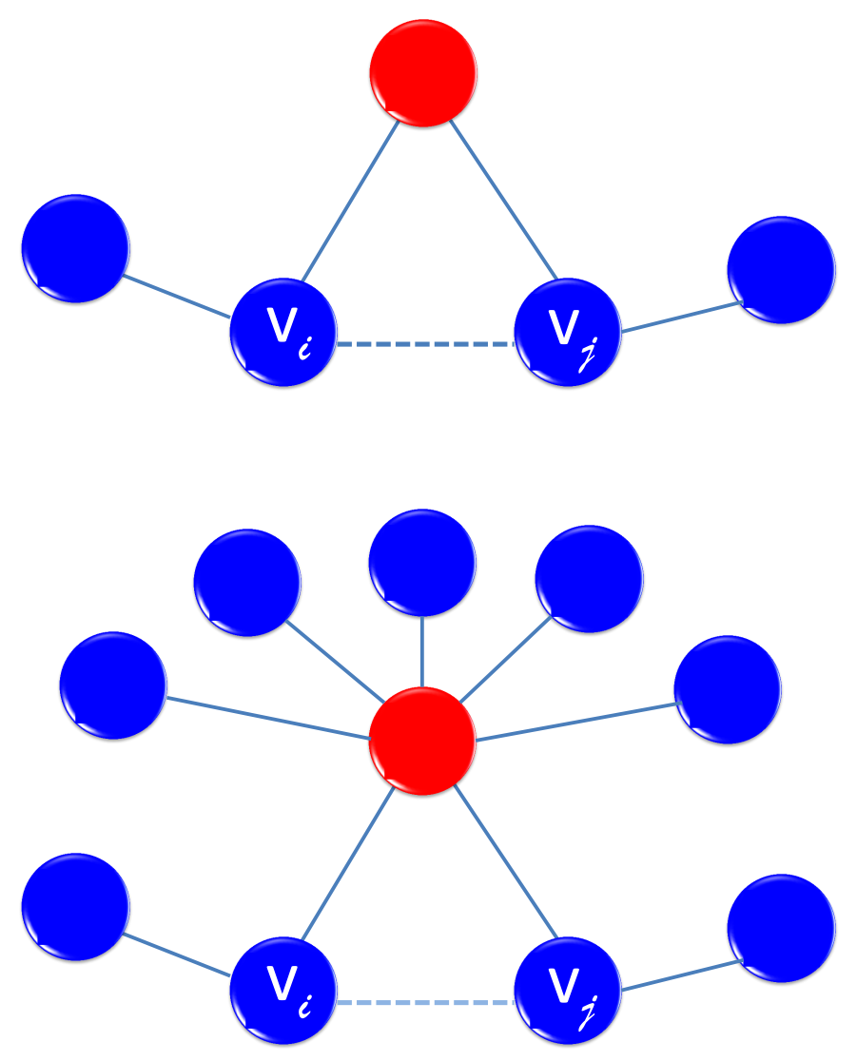
Large, sparse network have a large class imbalance, with potential links (Negatives) $>>$ new links (Positives).

Signal detection

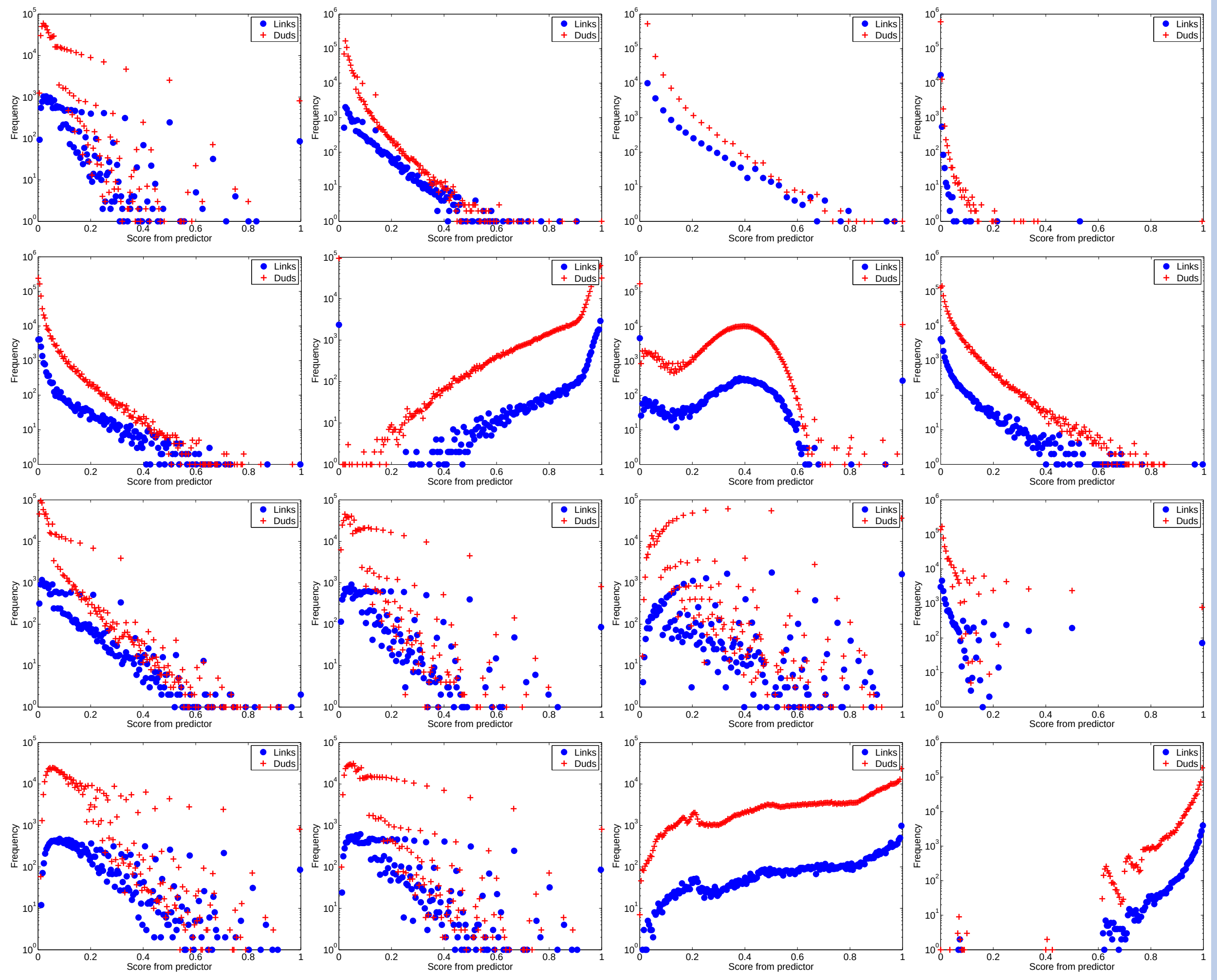
A large percentage ($\approx 35\%$) of new links that arise in the $t+1$ time step occur between users connected by chemical distance of $\delta = 2$ at time t. As such, we focus our attention solely on predicting triadic closure. Similarity indicices, such as Adamic-Adar

$$A(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{log(|\Gamma(z)|)}$$

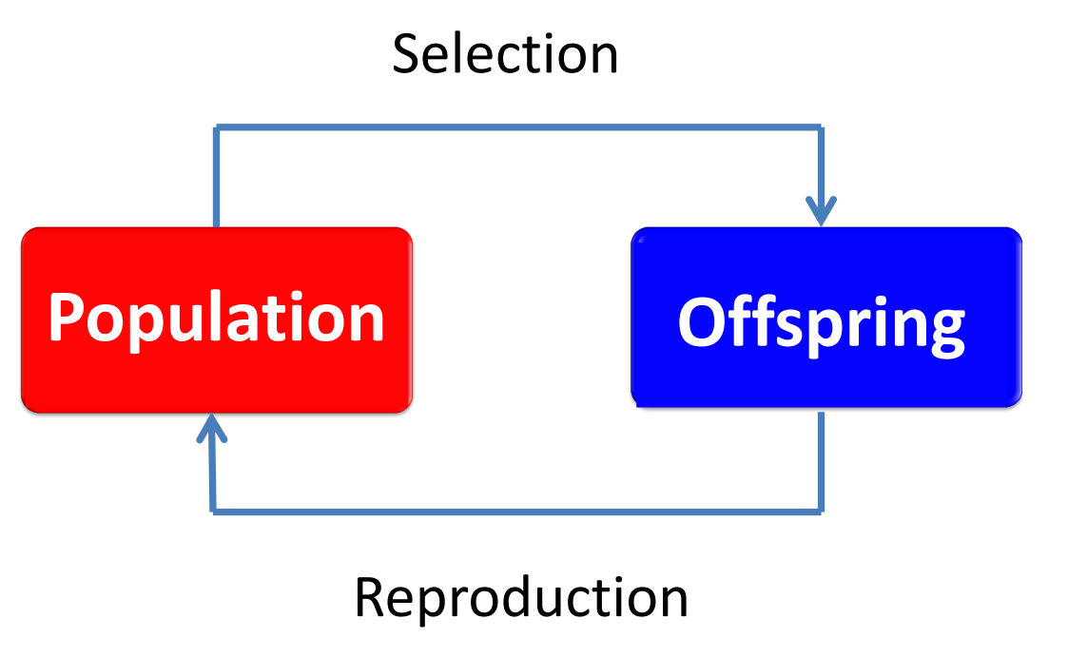
quantify the similarity of two nodes based on topological features. A weights rarer features more heavily and predicts a link in the bottom right plot as more likely than in the top right.



We compute 12 topological and 4 user-specific similarity indices for nodes with $\delta = 2$ at time t. Below: Scores for user-user pairs which exhibit a link (blue) are difficult to separate from pairs which do not (red) for the indices used in isolation (from L to R, Row 1 to 4: Jacard, Adamic-Adar, Common neighbors, Paths, Katz, Happsim, Wordsim, Preferential Attachment, Resource Allocation, Hup depressed, Hub promoted, LHN, Salton, Sorenson, Idsim, Tweetcountsim)

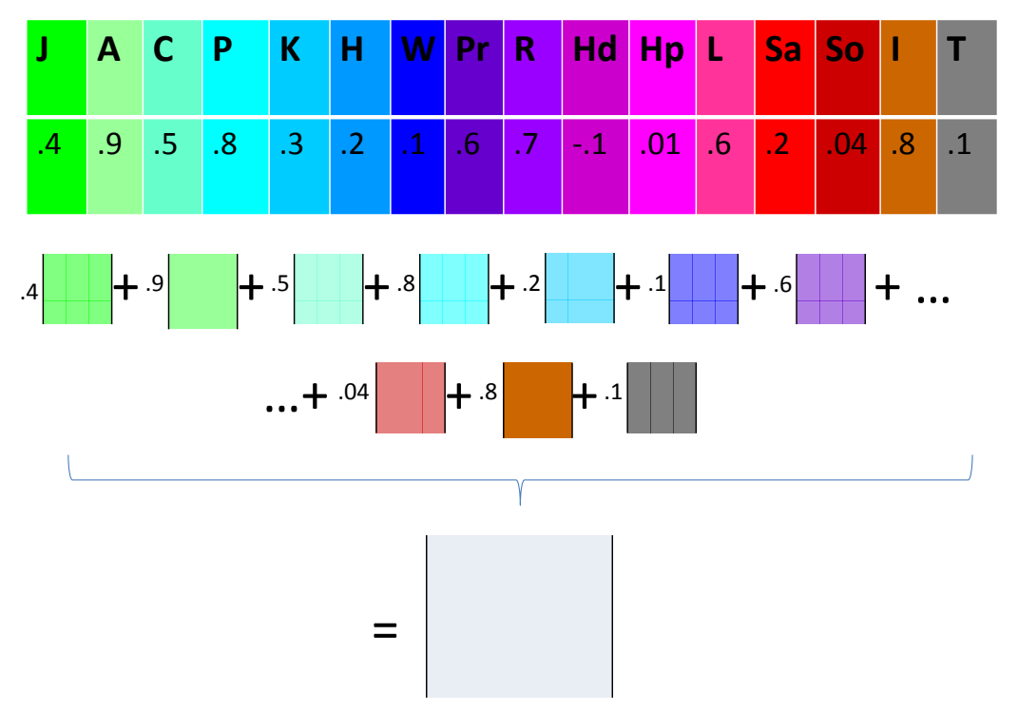


Covariance Matrix Evolutionary Stategy (CMA-ES)



At each generation, a multi-variate Gaussian cloud of candidate solutions (e.g., population) is generated in accordance with the covariance matrix. User-user pairs with the top$N$ scores from score $= \sum_{i=1}^{16} w_i S_i$ are those for whom a new link is predicted.

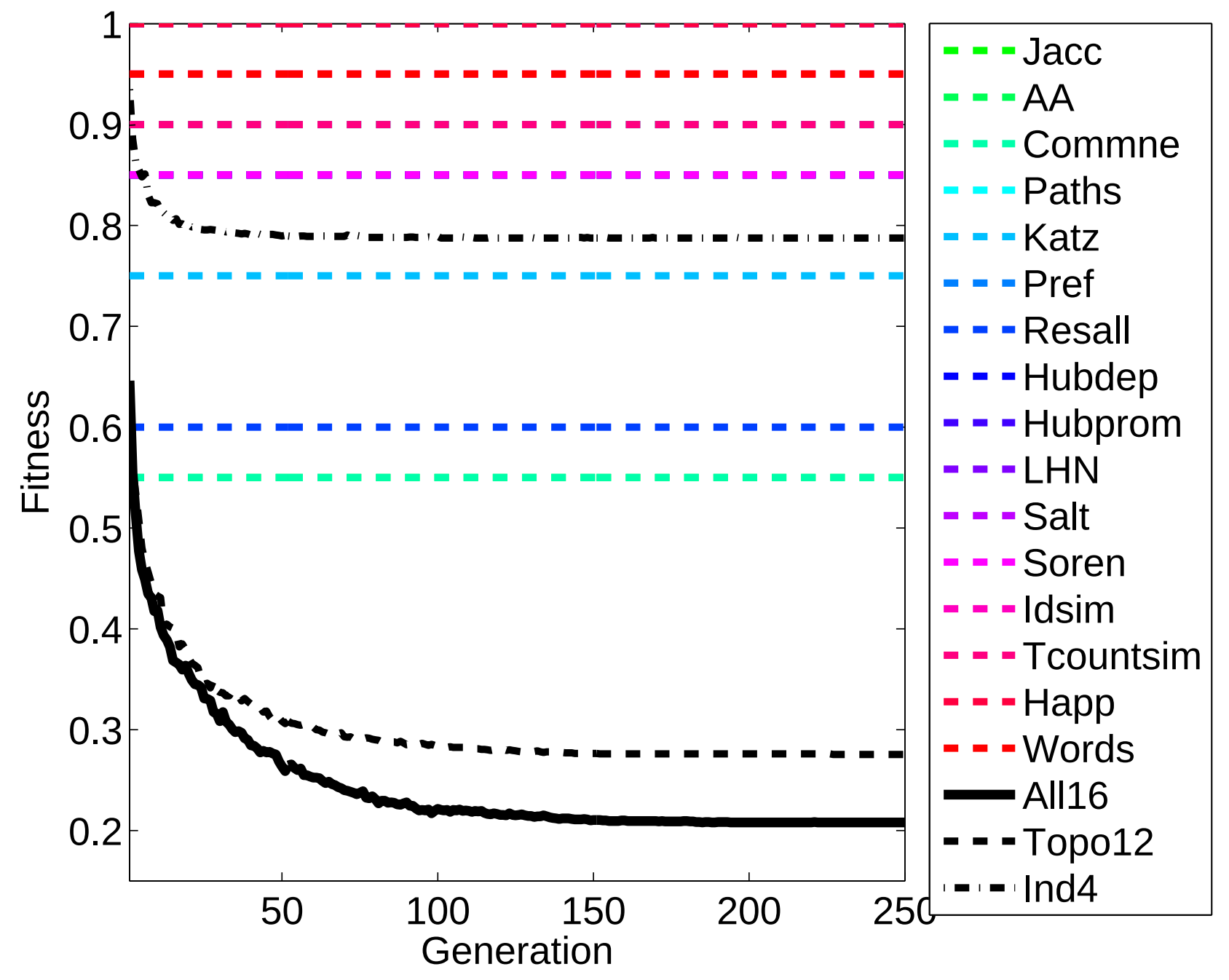
Covariance Matrix Evolutionary Strategy (CMA-ES) evolves candidate solutions to the link prediction problem. We divide our data into a training set (Weeks 1 through 6) and a validation set (Weeks 7 through 12). Our evolutionary algorithm begins with $\vec{w} \in \mathbb{R}^{16}$, initialized with real valued entries between 0 and 1.



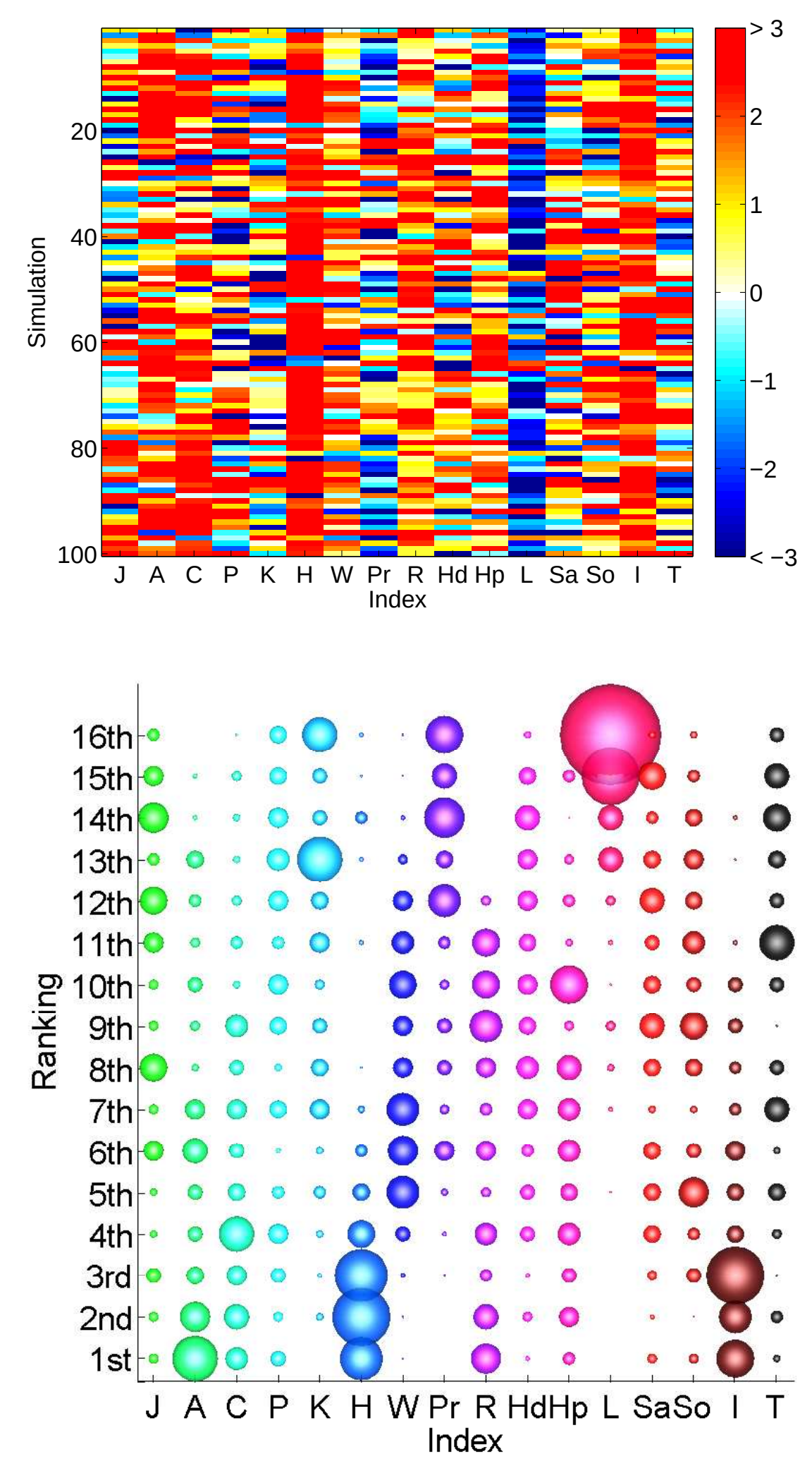
J	A	C	P	K	H	W	Pr	R	Hd	Hp	L	Sa	So	I	T
.4	.9	.5	.8	.3	.2	.1	.6	.7	-.1	.01	.6	.2	.04	.8	.1

Fitness

Selection operates on $\vec{w}$, whereby fitness is assessed as the proportion of the top$N$ links incorrectly predicted. As CMA-ES works to minimize fitness, the fitness plot (right) shows that the "all16" and "topo12" evolved predictors outperform all other similarity indices used in isolation.



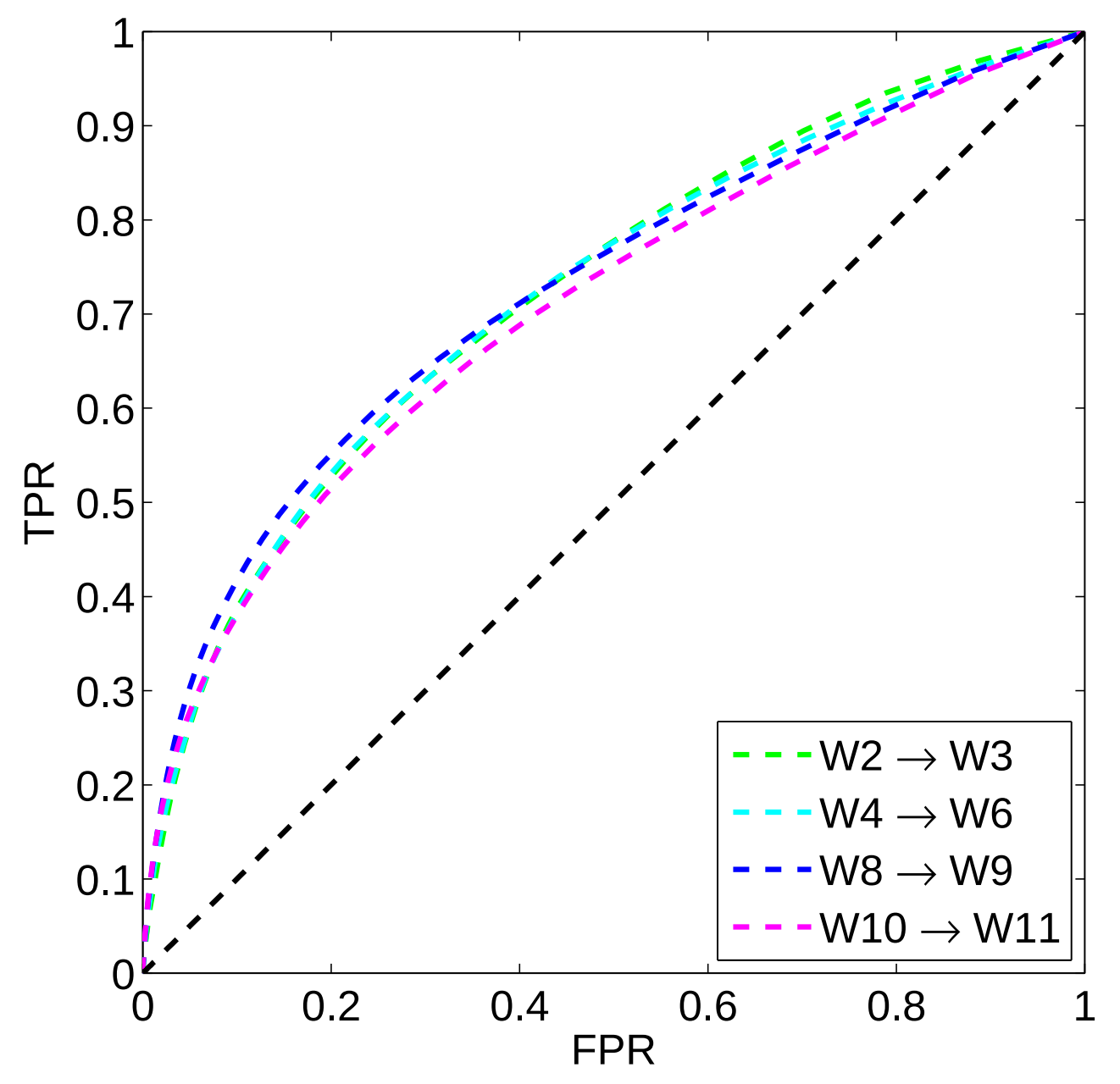
Best solutions



Top left: The 100 best candidate solutions which evolved after 250 generations of CMA-ES, $\vec{w}$, are shown as horizontal rows. The ith column signifies the w_i coefficient used in the linear combination of the weights. The color axis reveals the value of ith coefficient. Although there is considerable variability, user-user pairs which had high scores for the indices which evolve positive weights (e.g., Adamic-Adar, Common neighbors, Resource Allocation, Happiness and Id similarity) and low scores for the indices which evolve negative weights (e.g. Leicht-Holme Newman) were more likely to exhibit a future link. **Bottom left**: Frequency plot shows that Adamic-Adar, Common neighbors, Resource Allocation, Happiness and id similarity often evolved with the largest positive weights, while LHN often evolved to have the largest negative weight.

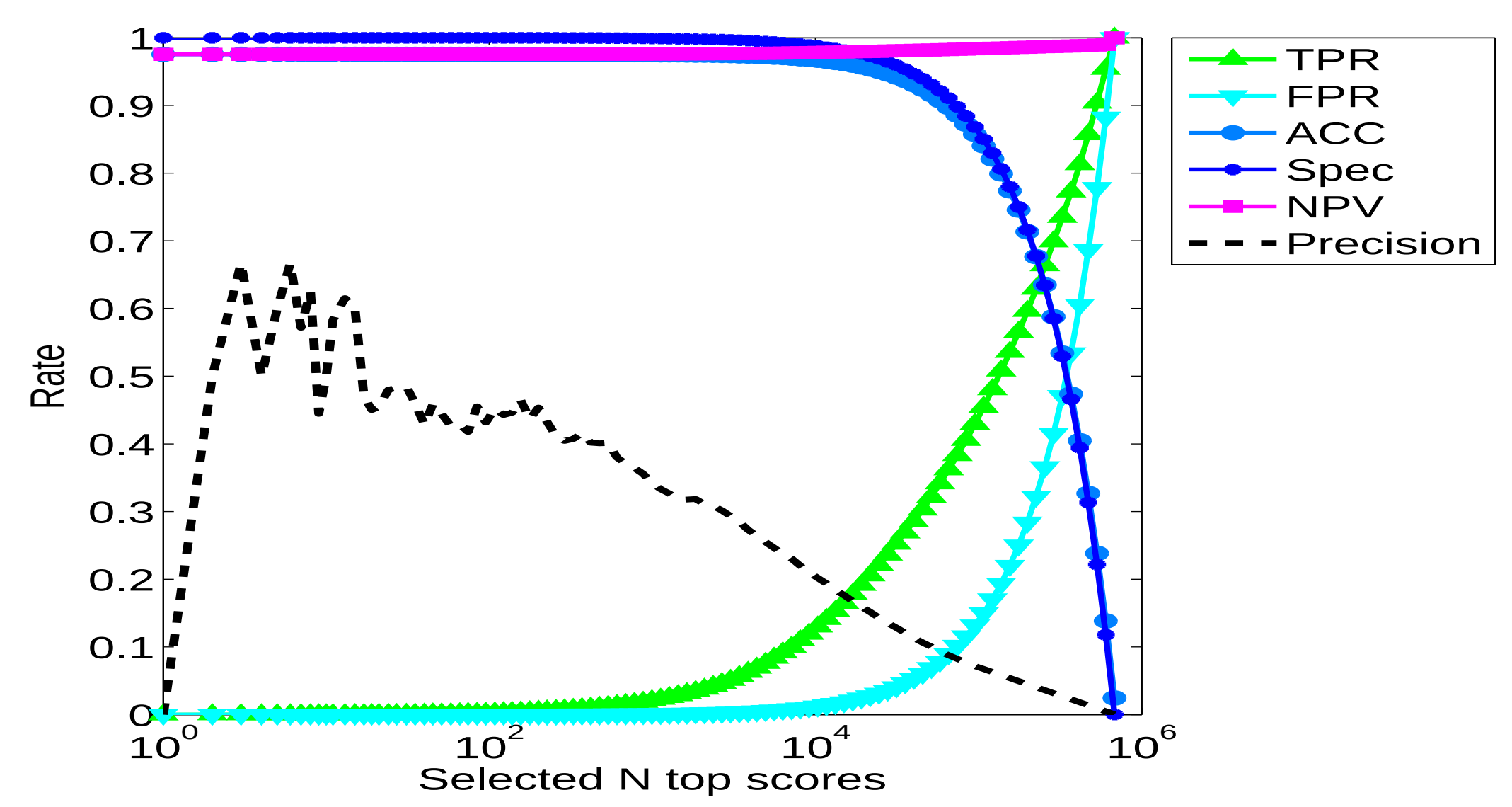
Receiver Operating Curve (ROC)

The Receiver Operating Characteristic (ROC) curve depicts the true positive rate (TPR) as a function of the false positive rate (FPR). A classification method which randomly assigns true or false to the presence of future links would, on average, have TPR equal to FPR. Successful classifiers have $TPR > FPR$ and this is often quantified by estimating the area under the curve (AUC) for the ROC. The AUC approximates the probability that a link predictor will assign a higher score to user-user pairs who exhibit a link in the next timestep than to user-user pairs who do not exhibit a link in the next timestep.



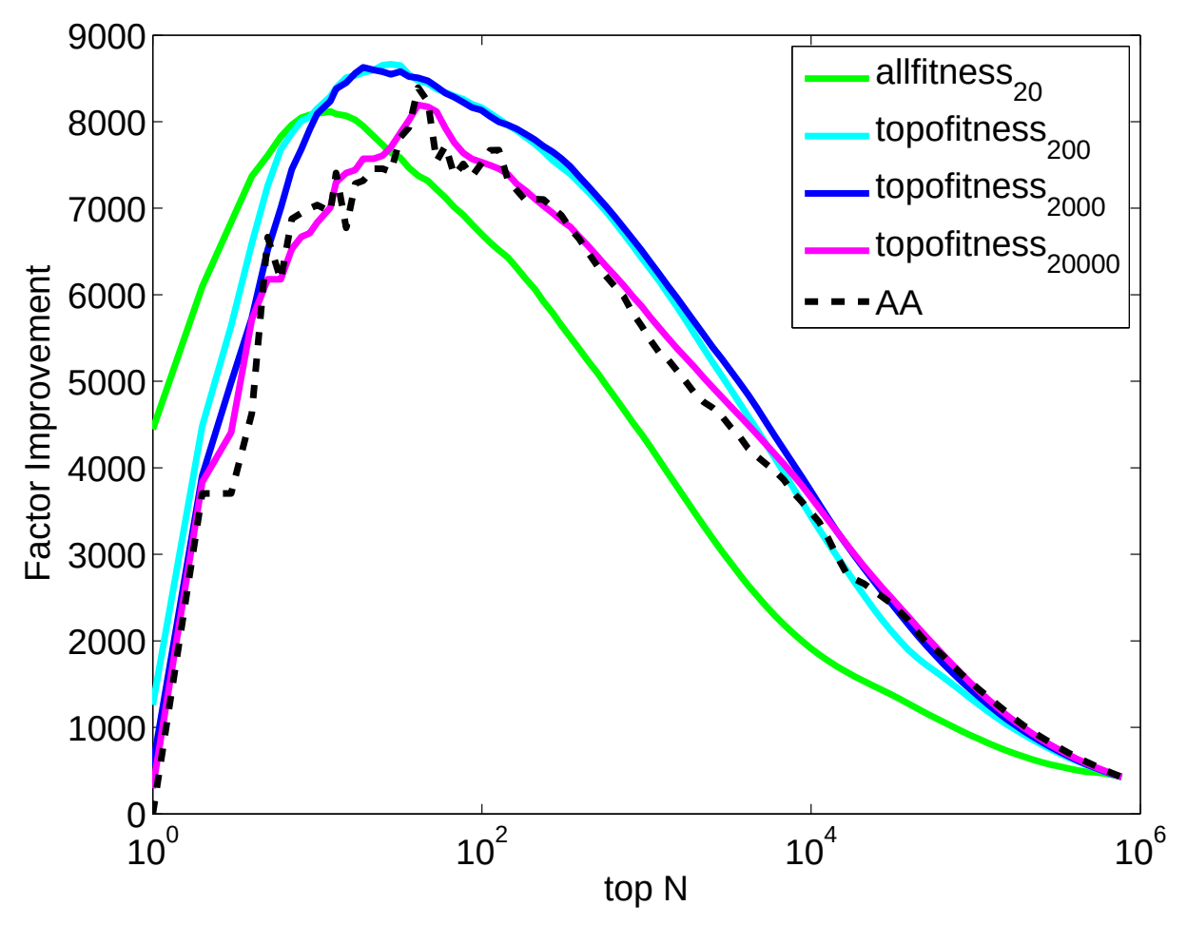
assign a higher score to user-user pairs who exhibit a link in the next timestep than to user-user pairs who do not exhibit a link in the next timestep. Above: $AUC > 0.70$ for all weeks.

Precision



Precision (black) is quite high for top$N$ link prediction on the order 20. Negative predictive value is consistently high (magenta) because of the large class imbalance. Specificity and Accuracy are also quite high for several orders of magnitude. We explore the potential impact of missing tweets on our predictor. We randomly select 50% of our observed tweets and construct the reciprocal reply subnetworks for Weeks 1 through 12. We identify the percent of links which are incorrectly labeled as false positives in the subnetworks because they are true positives in the larger, observed networks. Upon doing so, we see that precision is higher than reported for the subnetworks and suspect a similar trend for our observed data (above).

Factor Improvement



Notable works in the area of link prediction have reported the factor improvement over random link prediction. We follow suit and compute the factor improvement of our predictor over a randomly chosen pair of users. The probability of a randomly chosen pair of users who are not connected in week i to become connected in week $i+1$ is $\frac{|\text{Edges}_{\text{new}}|}{\binom{|V(G)|}{2} - |\text{Edges}_{\text{old}}|}$. There are 44,439 nodes in the validation set and, as a sample calculation, $|E(G_{\text{validate}})| = 71,927$ in week 7. There are 53,722 new links that occur from week 7 to 8. Thus, the probability of a randomly chosen pair of nodes from Week 7 exhibiting a link in Week 8 is approximately $\frac{53,722}{\binom{44,439}{2} - 71,927} \approx .0054\%$. We observe significant factors of improvement over randomly selected new links, usually on the order of 10^3

Acknowledgments

The authors wish to acknowledge the Vermont Advanced Computing Core, which is supported by NASA (NNX-08AO96G) at the University of Vermont which provided High Performance Computing resources that contributed to the research results reported within this poster. CAB was supported by the UVM Complex Systems Center Fellowship Award, PSD was supported by NSF Career Award # 0846668. CMD, PSD, and MRF were also supported by a grant from the MITRE Corporation.

References

Bliss, Catherine A., Kloumann, Isabel M., Harris, Kameron Decker, Danforth, Christopher M. and Peter Sheridan Dodds. (2012). Twitter reciprocal reply networks exhibit assortativity with respect to happiness, textitJournal of Computational Science 3:388-397.

Dodds, Peter Sheridan, Harris, Kameron Decker, Kloumann, Isabel M., Bliss, Catherine A. and Christopher M. Danforth. (2011). Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter, *PLoS ONE* 6(12):e26752.

Liben-Nowell, Kleinberg. (2007). The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology* 58(7):1019-1031.