Suspicion scoring based on guilt-by-association, collective inference, and focused data access[1]

Sofus A. Macskassy and Foster Provost
New York University
44 W. 4th Street, New York, NY 10012

{smacskas,fprovost}@stern.nyu.edu

Keywords: Predictive analysis, Link analysis, Social network analysis, Counter-terrorism

Abstract

We describe a guilt-by-association system that can be used to rank entities by their suspiciousness. We demonstrate the algorithm on a suite of data sets generated by a terrorist-world simulator developed under a DoD program. The data sets consist of thousands of people and some known links between them. We show that the system ranks truly malicious individuals highly, even if only relatively few are known to be malicious ex ante. When used as a tool for identifying promising data-gathering opportunities, the system focuses on gathering more information about the most suspicious people and thereby increases the density of linkage in appropriate parts of the network. We assess performance under conditions of noisy prior knowledge (score quality varies by data set under moderate noise), and whether augmenting the network with prior scores based on profiling information improves the scoring (it doesn't). Although the level of performance reported here would not support direct action on all data sets, it does recommend the consideration of network-scoring techniques as a new source of evidence in decision making. For example, the system can operate on networks far larger and more complex than could be processed by a human analyst.

1. Introduction

This paper studies *suspicion scoring*: ranking individuals by their estimated likelihood of being malicious. In particular, we address suspicion scoring in networks of people, linked by communications, meetings, or other associations (e.g., being in the same vicinity at the same time). Our system makes use of the simple-yet-ubiquitous principle of homophily [Blau 1977; McPherson et al. 2001]; social research has shown repeatedly that a person is more likely to associate with people who share similar interests or characteristics. Homophily is the basis of a simple guilt-by-association algorithm: estimate suspicion level by counting malicious associates.

Suspicion scoring based on networked data has been used successfully, although typically in an ad hoc manner, for fraud detection. The "dialed digits" monitors discussed by Fawcett and Provost score an account highly if it calls the same numbers called by known fraudulent accounts [Fawcett and Provost 1997]; the "communities of interest" of Cortes et al. explicitly represent the network neighborhoods around telephone accounts as a basis for suspicion scoring [Cortes et al. 2001]. We extend such methods by propagating suspicion through the association network, and conducting suspicion-based acquisition of additional data.

One problem with using the simple homophily-based guilt-by-association algorithm in large networks is that few people may be known to be malicious. Often none of an individual's associates are known to be either malicious or benign. However, if the association graph is well connected, then following linkages of associations is likely eventually to lead to at least one individual who is known or strongly suspected to be malicious. Based on this idea, we overcome the problem of sparse knowledge by propagating suspicion scores through the association network until all suspicion scores stabilize. In particular, we use an adaptation of the relaxation labeling method shown to yield good performance for hypertext classification by Chakrabarti et al. [1998].[2]

Relaxation labeling works well if the association graph is well-connected. For intelligence data, one must consider the difference between the *true* association network and the network of *known* associations. The association network may be known only partially. We address this via suspicion-based data acquisition, using current suspicion scores to acquire additional connections to improve the suspicion propagation. In a realistic setting, acquiring association links (involving subpoenas for transaction records, surveillance, interviews, phone taps, etc.) is costly in terms of money, resources, legal issues, and public perception. We attempt to minimize costs by acquiring such "secondary data" only for the people with the highest estimated suspiciousness. This heuristic works well in the data we have studied.

[1] To appear in the International Conference on Intelligence Analysis, 2005.

[2] This "relational neighbor" algorithm with belief propagation was introduced previously in a workshop paper [Macskassy and Provost 2003].

1) A *relational classifier* which generates a suspicion score for a particular individual, p_i, given the known associations of p_i and the strengths of those association links. 2) A *collective inference* technique to propagate scores throughout the network. 3) An adaptive technique for *acquiring data* to increase the density of connections in the network.

Table 1: Guilt-by-association main components.

1) Acquire information on all people initially known to be malicious. 2) Generate suspicion scores for all individuals with unknown scores. 3) Get information on the top k individuals not yet queried (k = 50 for this paper). 4) Generate new suspicion scores. 5) Repeat steps 3 and 4 until some stopping criterion is met (in this paper: either when all individuals in the data set have been queried against or when we reach the 25th iteration.)

Table 2: Data Acquisition algorithm.

2. Guilt-by-association, Collective inference, and data acquisition

Our scoring algorithm consists of three main components listed in Table 1. The first two components are part of a network learning toolkit called NetKit-SRL [Macskassy & Provost. 2004]. This open-source toolkit, written in Java 1.5, is publicly available and contains methods for learning patterns more complicated than simple guilt-by-association. The third component is a data acquisition wrapper which uses this toolkit in its inner loop.

2.1 Relational Classifier

The relational classifier used in the study is a simple "relational neighbor" model, based on the principle of homophily and a first-order Markov assumption [Macskassy and Provost 2003, 2004]. The model estimates suspicion as the weighted sum of the suspicions of the immediate neighbors in the association network. Specifically, the score of person *i* is:

$$s(p_i) = \frac{1}{Z} \sum_{p_j \in N_i} w_{i,j} \cdot s(p_j)$$

where N_i is the set of known associates of person p_i and $w_{i,j}$ is the strength of the association between persons *i* and *j*—in our application defined as the number of times p_i and p_j have been known to interact. The score, $s(p_j)$, is the current suspicion score of person p_j (note the similarity of our method, paired with the updating method described below, to Hopfield Networks [Hopfield 1982] and Boltzmann machines [Ackley et al. 1985]). For people whose status is known (good or bad), this is static—viz., 1 for 'bad' and 0 for 'good'. Z is the sum of weights $w_{i,j}$, to keep all scores between 0 and 1.

2.2 Collective Inference

When only a few malicious individuals are known, there will be neighbors who (initially) have no value for $s(p_j)$. To deal with this scenario, first recognize that if we had estimates of the unknown scores, then we could apply the relational classifier to estimate $s(p_i)$. Second, the scores of p_i and p_j are clearly interrelated and estimating one will have an influence the other. We therefore estimate all unknown scores simultaneously or "collectively" [Jensen et al., 2004]. As it is not tractable to perform exact inference to estimate the full joint probability distribution over a large network, we use an approximation technique. In particular, we use an adaptation of relaxation labeling, based on the work of Chakrabarti et al. [1998]. Relaxation labeling "freezes" the current estimated scores and then updates all estimates pseudo-simultaneously to generate new estimates. It does so repeatedly until the estimates converge. Unfortunately, this often leads to oscillation between two distinct sets of world-estimates. Therefore, we apply simulated annealing to enforce convergence. More formally:

$$s(p_i)^{(t+1)} = \alpha^{(t+1)} * s(p_i)^{(t)} + (1-\alpha^{(t+1)}) * \left(\frac{1}{Z} \sum_{p_j \in N_i} w_{i,j} \cdot s(p_j)^{(t)} \right)$$

where t is the iteration step and $\alpha^{(t)}$ the temperature, with

$$\alpha^{(0)} = \boldsymbol{c}$$
$$\alpha^{(t+1)} = \beta * \alpha^{(t)},$$

where $\boldsymbol{c}$ is a starting constant and β is a decay constant. We use the values 1 and 0.99 for $\boldsymbol{c}$ and β, respectively, and stop after 100 iterations.

Relaxation labeling and other collective inference techniques require initial estimates to bootstrap the inference. We initialize scores to 1 for initially "known" malicious people (and freeze them) and 0.01 for the rest. If we had had knowledge of benign people, we would have initialized those scores to 0.

2.3 Data Acquisition

As discussed above, it may be possible (at a cost) to augment the association network incrementally. The strategy used in this paper is shown in Table 2, which acquires additional information (associations and possibly unknown associates) about the most suspicious people.

3. Case Study

Using simulated data, we evaluate whether this method can produce accurate rankings of individuals by suspicion scoring. Specifically: are the highest-scoring individuals predominantly malicious? Our study is fourfold,

Data set	True size	Number bad	Initial size	True bad	False bad	Noise
5046	13236	1484	4212	143	52	0.267
5048	13756	2173	9601	226	0	0
5049	988	269	766	62	0	0
5050	1008	316	439	103	336	0.765
5052	986	278	292	82	210	0.719
5053	1002	274	332	116	216	0.651
5056	1022	300	317	99	218	0.688
5062	9897	2852	3745	500	0	0
5063	9998	2823	4825	828	276	0.25
5065	16046	7574	5907	1264	82	0.061
5066	16743	8002	5332	1284	173	0.119

Table 3: Characteristics of Synthetic Data sets. True size refers to the number of individuals in the true synthetic world and **Number bad** refers to the total number of truly malicious individuals. **Initial size** refers to the number of individuals included in the primary data; **True bad** refers the number of individuals tagged as 'bad' who truly were bad in the world and the **False bad** refers to the number of individuals falsely tagged as 'bad'. The error rate (**Noise**) of these labelings ranges from none (0) to very high as seen in the 505x data sets. Note that step 1 in Table 2 above must query the secondary database for information on all 'false bad' as well as all 'true bad' individuals.

assessing: (1) the initial rankings; (2) the improvement as we acquire more information, (3) how good the initial knowledge of maliciousness must be (i.e., how much noise can be tolerated), and (4) whether adding profiling-based initial scores improves the final scoring.

3.1. Data

There are many varieties of intelligence data—no single comparison of classified and synthetic data will be comprehensive. The data we use for this paper were generated by a flexible simulator as part of a DoD program to assess the feasibility of large-scale information systems to help identify terrorists. The synthetic data generated by this simulator are moderately sized examples of structured data representing terrorists and benign entities who are conducting activities over an extended period of time. The data are contained wholly in a single data source and are self-consistent, neither of which is reliably true of classified data. However, the data do replicate a range of noisy and poorly observed activities, and the entities are intentionally obscured to simulate lack of knowledge, obfuscation, poor data-entry practices, multiple identities, etc. Nonetheless, we do not claim that the data fully replicate the limitations of actual data collection, aggregation and enrichment that the intelligence community routinely experiences. We use data sets generated for the purposes of DoD program evaluation. We do not create data sets ourselves for this paper.

One run of the simulator generates three databases:

1) "primary" data that are known ex ante. These often are sparse and may contain partial (or no) information on any particular individual or group;
2) "secondary" data consisting of information which can only be acquired by querying (theoretically at a cost) to get information on a particular individual;
3) "truth" data, for evaluation, consisting of what really happened in the world.

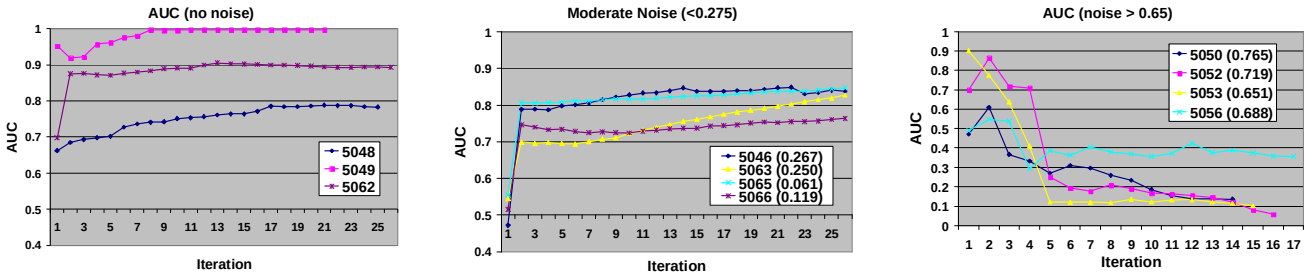
The first two databases together reflect what possibly can be observed. They are potentially corrupt and contain only a subset of the complete truth. Further, the data never give hard evidence that an individual is benign, and therefore we only "know" about some malicious individuals—those who are known to belong to one or more terrorist groups. Sometimes this knowledge is wrong. We evaluate the suspicion scoring on eleven data sets, whose characteristics are shown in Table 3.

3.2 Results

We evaluate the suspicion scoring using the Area Under the ROC Curve (AUC), which is equivalent to the Mann-Whitney-Wilcoxon statistic and computes the probability that a randomly selected malicious person would be given a higher suspicion score than a randomly selected benign person. Therefore, an AUC of 0.5 means that a scoring is no better than random guessing (the ranking is well shuffled); a value of 1 indicates a perfect ranking—all the malicious people get higher scores than all the benign people. Since we cannot generate scores for unseen individuals or for individuals with no known associations, these are not considered when calculating the AUC. In order to address how noise affects performance, we group the data sets into three categories: no noise (5048, 5049, 5062), low to moderate noise (5046, 5063, 5065, 5066), and very high noise (5050, 5052, 5053, 5056).

Figures 1(a)-(c) show, for each category, how the system performed throughout its data acquisition run. Iteration 1 represents using only initially known data and iteration 2 is the performance after querying all individuals whose suspicion score was known initially. Figure 1(a) shows a wide range of performances from almost perfect (5049) to just below AUC=0.8 (5048). In all three cases, we see that performance increases as we gather more data. We also see that guilt-by-association is able to perform much better than random ranking (AUC = 0.5).

Figure 1(b) shows the performance of the suspicion scores on the moderate-noise data sets. As we can see, using only primary data generally results in performance no better than random. However, as we query the secon-



(a) no noise (b) moderate noise (c) very high noise

Figure 1: AUCs using active data acquisition.

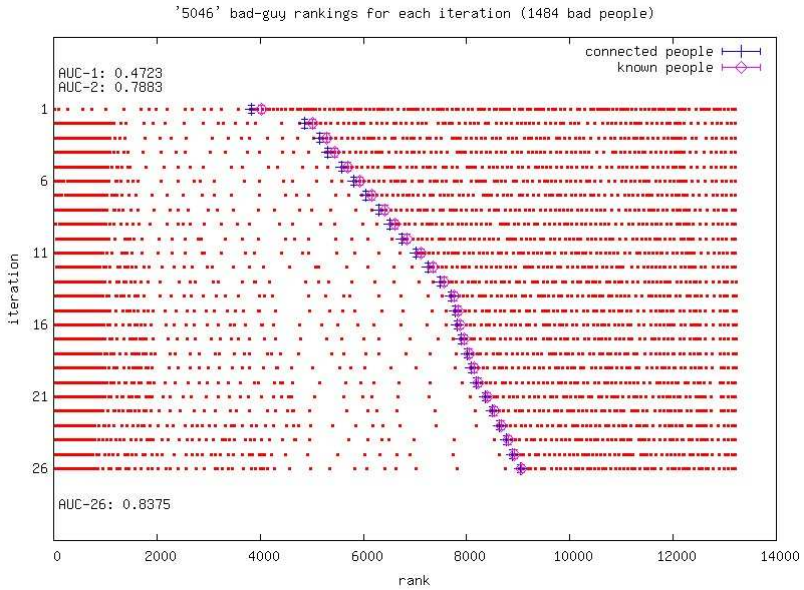


Figure 2: Ranking of bad guys after each iteration of data acquisition on data set 5046 (moderate noise). Iteration 1 corresponds to using initially known data, and iteration 2 is after querying on all individuals with known suspicion scores. A red dot represents the rank of one true bad guy, where lower rank means having a higher score. Red dots at the far left correspond to bad guys in the top-ranks. The figure shows how many individuals have been seen with respect to the true world (purple diamond) as well as how many of those have known interactions with other people (blue cross).

dary data the performance quickly improves to achieve AUC values around 0.8. It is noteworthy that the system overcomes the initial noisy labels, showing that it is robust to even moderate noise where 1 out of 4 people were mistakenly judged to be malicious. This is because "bad" people in these data communicate just as much as "good" people, but to far fewer people. Hence their associative strength to other "bad" people is much stronger.

Figure 1(c) shows the performance of the suspicion scoring on the high-noise data. In this case, more than half of the individuals tagged as "bad" are actually "good". Even the scores that start off well deteriorate quickly, and all end up with ranking much worse than random (note the different vertical scale), because the algorithm is actually propagating knowledge of goodness; were we simply to flip the scores, then the system would perform quite well in 3 of the 4 cases!

Figure 1 only tells part of the story. For an analyst, knowing that the system can achieve an AUC of 0.8 does not necessarily mean that the system is useful. Although the system, in general, will rank suspicious people higher, when considering 10000 people, the top 100 could potentially be primarily good with the next 900 being primarily bad. Albeit unusual, this would achieve a relatively high AUC, but not be very useful for analysts who only have time to look at a select few individuals.

Usually for rankings such as suspicion scorings, the density of entities of interest is highest at the very top of the list, especially if the scores are estimated probabilities of membership in a class (e.g., malicious individual), and so the top of the list contains the individuals with the highest estimated probabilities. For a given AUC value, how dense one expects the top of a ranking to be depends primarily on the marginal probability of entities of interest in the data (this and related issues are treated in detail elsewhere [Provost and Fawcett, 2001]). An AUC of 0.8 may have 99 truly malicious individuals in the top 100 highest-suspicion individuals, or it may have 10. In either case, the system may be useful to an analyst, depending on the application and how the ranking will be

	Data	Iteration 1	Iteration 2	Last Iteration
No Noise	5048	86	98	100
	5049	59	78	100
	5062	100	96	100
Moderate Noise	5046	21	51	99
	5063	56	42	34
	5065	93	80	84
	5066	55	77	84
Large Noise	5050	23	44	5
	5052	32	45	1
	5053	21	46	4
	5056	22	36	18

Table 4: How many truly bad people were in the top 100 after iteration 1 (using only "primary" data), iteration 2 (after querying individuals "known" to be malicious), and after the last iteration.

used (e.g., as a primary basis for action versus as an alternative source of evidence to augment existing practices).

To illustrate the effectiveness of the scorings for these data sets for a particular threshold, we analyze the number of truly malicious individual at the top of the suspicion rankings. Figure 2 shows a plot of data set 5046 (13236 people, 1484 of them being truly malicious, moderate noise), where each line in the graph represents one iteration of the data acquisition, and each (red) dot represents a malicious individual in the ranking.

Figure 2 shows qualitatively what is happening after each iteration, where the "malicious" people clearly are being ranked higher and higher as a group. If we look across the 11 data sets, we can ask how many truly malicious people are among the top 100 most suspicious. We focus on three points of interest: initial data before any acquisition (Iteration 1), performance after querying all individuals "known" to be malicious (Iteration 2), and after the data acquisition is done (Last iteration). Table 4 shows the results for the 11 data sets, grouped by their noise level. Table 4 shows quantitatively what Figure 1 was telling us: In the no-noise group, we get very good rankings even using only primary data, where the increasing AUC score reflects what happens below the top 100. The moderate noise group shows mixed results, where 2 of the data sets (5046 and 5066) greatly increase the density of malicious individuals whereas 5065 is relatively stable and 5063 shows a decrease. Again, the lift in AUC values reflects what happens below the top 100. Finally, we see in the very large noise group that by the final iteration the system has almost no malicious individuals in the top 100. Again, remember that the absolute numbers (e.g., "precision" of 84 out of 100) reflect the marginal probability of being malicious in a particular data set; the ROC curve is independent of this probability, which is one reason why 5063 and 5046, although having very similar ranking ability (AUC) have very different precision for a fixed threshold [Provost and Fawcett, 2001].

3.3 Adding profiling information

For the results presented so far, we have assumed that the only information available ex ante is the set of possibly noisy labels for a subset of the nodes in the network. The suspicion scores were determined solely by the quality of these labels and the structure of the association network. However, additional information may be available for suspicion scoring. There may be suspicion models based on characteristics of individuals, or other less-certain prior knowledge. Such additional knowledge can augment the guilt-by-association suspicion scoring in several ways. The most straightforward is to use it to set the prior suspicion scores for the network. We model such additional knowledge simply by conditional score distributions, e.g., one for truly malicious individuals and one for benign individuals. Under such a framework, the prior scenario corresponds to: a uniform distribution (score=1) for the "known" individuals and a uniform unconditional distribution (score=0.01) for the rest. Statistical profiling of individuals based on their characteristics could be modeled as a pair of class-conditional distributions, benign versus malicious. Profiling is effective to the extent that the class-conditional distributions are separated.

It is intuitive that augmenting a network with very high-quality knowledge should improve suspicion scoring. However, it is not clear whether adding lower-quality knowledge should improve the scores, or whether the propagation of the "known labels" is sufficient. We augmented the simulated domains by adding profiling scores based on separated class-conditional distributions. Specifically, we generated scores for truly malicious (benign) individuals by randomly sampling from a normal distribution with a mean of 0.55 (0.45) and standard deviation of 0.25. This resulted in a slight, but not noticeable, improvement in three of the 11 data sets (5046, 5048 and 5062). In none of the cases did the priors hurt performance. Increasing the separation and decreasing the standard deviation increased the improvement for those three data sets and surprisingly had no discernable difference for the remaining 8. Figure 3 shows the most noticeable improvements when using distributions with means 0.9 (malicious) and 0.1 (benign) and a standard deviation of 0.05 (essentially perfect separation).

We examined the data sets to investigate whether any characteristics explain why the augmented priors helped in only 3 of the 11 data sets. The only seemingly explanatory factor was the fraction of "known" individuals. The three data sets where the augmented priors helped were also the three data sets that had the smallest ratio of known individuals to total individuals. Further investigation reveals that even perfect initial profiling scores get "washed out" by the static labels and their propagation through the network. In some sense, having only few labels "helped" in that the initial priors were not entirely dominated by the known labels. This is an important limitation which requires further research.

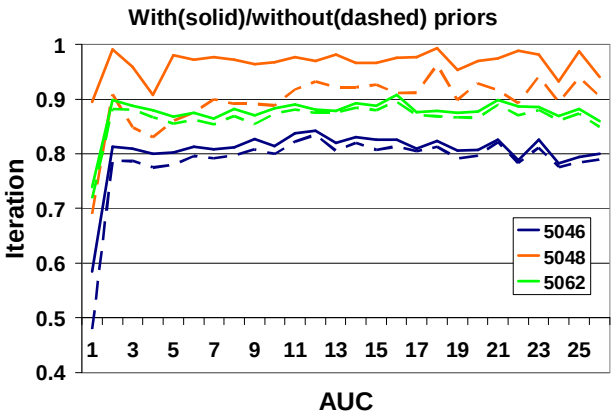


Figure 3: Improvements in performance when augmenting prior-scores by randomly drawing from very well separated class-conditional distributions.

4. Other limitations and issues

The system we described here has other notable limitations. First, it relies on the network and will therefore suffer greatly if there is not enough associative information in the graph. Second, while it can be relatively robust to moderate noise, it can perform poorly if the initial knowledge is bad. Third, as just mentioned, we found that the static labels and the network structure would dominate the final scores to the point where initial priors had no effect. We need to understand better how profiling scores can help the overall performance. This is an issue that is likely to impact many collective inference techniques and needs to receive more attention.

5. Final Remarks

We described and evaluated a guilt-by-association system for generating suspicion scores based on entities' known associates. The system is notable for several reasons. First, it is able to generate remarkably good rankings even when very few individuals have known suspicion scores. Second, it can be relatively robust even to moderate noise in initial scores. Third, it works remarkably well considering that it only uses suspicion scores and the network, but no profiling. Finally, it can be used as a data gathering tool not only to perform focused data acquisition of suspicious people, but also to further improve its ranking—and in the process often learn/acquire data about suspicious people that were not initially in the database.

We evaluated this system on a range of data sets, varying in both size and noise level. We saw that the system was robust to moderate noise but failed when the majority of the initial scores were wrong. We further showed that it was possible to improve rankings by using a focused data acquisition technique, sometimes being able to achieve almost perfect separation between malicious and benign people in the network.

Lastly, we did a preliminary investigation into augmenting the scoring with other uncertain-but-better-than-random knowledge (as from a profiling system). We found that priors had little-to-no effect due to the dominance of the scores propagated from the static labels. This is a problem which can have an impact on many collective inference techniques. An important open question is how one should combine relational and local information properly such that one does not dominate the other.

Acknowledgments

This research was sponsored in part by the Air Force Research Laboratory, Air Force Materiel Command, USAF, under Agreement number F30602-01-2-0585. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government. This work was also funded in part by a grant from the New York Software Industry Association.

References

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski (1985) A learning algorithm for Boltzmann machines. *Cognitive Science, 9,* 147-169, 1985.

P. M. Blau (1977) Inequality and Heterogeneity: A Primitive Theory of Social Structure. New York: Free Press, 1977.

S. Chakrabarti, B. Dom, and P. Indyk. (1998) Enhanced Hypertext Categorization Using Hyperlinks. In *ACM SIGMOD International Conference on Management of Data, 1998.*

C. Cortes, D. Pregibon and C. Volinsky (2001) Communities of Interest, The *Fourth International Symposium of Intelligent Data Analysis (IDA 2001)*, 2001.

T. Fawcett and F. Provost (1997) Adaptive Fraud Detection. *Data Mining and Knowledge Discovery, 3,* 291-316, 1997.

D. Jensen, J. Neville and B. Gallagher (2004) Why collective inference improves relational classification. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

J. J. Hopfield (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79,* 2554-2558, 1982.

S. A. Macskassy and F. Provost, (2003) A Simple Relational Classifier. In the *2nd Workshop on Multi-Relational Data Mining (MRDM) at KDD-2003.*

S. A. Macskassy and F. Provost, (2004) Classification in Networked Data: A toolkit and a univariate case study. *CeDER Working Paper #CeDER-04-08*, Stern School of Business, New York University, NY, NY 10012, 2004.

M. McPherson, L. Smith-Lovin, and J. M. Cook. (2001) Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology, 27,* 415-444, 2001.

Provost, F. and T. Fawcett. (2001) Robust Classification for Imprecise Environments. *Machine Learning 42,* 203-231, 2001.