# A Robust Collective Classification Approach to Trust Evaluation

Xi Wang, Mahsa Maghami, and Gita Sukthankar

Department of EECS
University of Central Florida
Orlando, FL
{xiwang,gitars}@eecs.ucf.edu,mmaghami@cs.ucf.edu

**Abstract.** In this paper, we present a collective classification approach for identifying untrustworthy individuals in multi-agent communities from a combination of observable features and network connections. Under the assumption that data are organized as independent and identically distributed (i.i.d.) samples, traditional classification is typically performed on each object independently, without considering the underlying network connecting the instances. In collective classification, a set of relational features, based on the connections between instances, is used to augment the feature vector used in classification. This approach can perform particularly well when the underlying data exhibits homophily, a propensity for similar items to be connected. We suggest that in many cases human communities exhibit homophily in trust levels since shared attitudes toward trust can facilitate the formation and maintenance of bonds, in the same way that other types of shared beliefs and value systems do. Hence, knowledge of an agent's connections provides a valuable cue that can assist in the identification of untrustworthy individuals who are misrepresenting themselves by modifying their observable information. This paper presents results that demonstrate that our proposed trust evaluation method is robust in cases where a large percentage of the individuals present misleading information.

**Keywords:** collective classification, homophily, agent reputation and trust

## 1 Introduction

Deciding whom to trust in the absence of direct transactional history is a difficult problem [35] for an individual agent interacting with an open system of self-interested agents. One oft-used mechanism is the direct solicitation of reputation information from a trusted source [36, 45], or multiple, less-reliable sources [13], to avoid deceptions perpetrated by groups of colluding agents. Yet, what if it is not possible to directly query an agent's reputation, either due to communication constraints or a lack of willingness from an agent's fellows to directly testify about past transactions? Here, we suggest that the structure of the network implicitly bears witness to the trustworthiness of the connected agents, regardless of whether the agents directly volunteer reputation information. Our collective classification framework for trust evaluation leverages a combination of observable features and network connectivity to improve performance over non-relational classification paradigms, in addition to making the trust evaluation process more robust against the deceptive efforts of untrustworthy agents.

In this paper, we describe collective classification and show how it can be used for general trust evaluation problems such as coalition building in social networks. Section 2 provides an overview of the social forces driving the creation of human networks and describe how the end result of these forces is an increase in the informative power of the network. Section 3 describes our specific agent reputation and trust scenario in which an individual agent has to evaluate the trustworthiness of a large number of surrounding agents. To make the collective classification process tractable for large datasets we use a local algorithm (Iterative Classification Algorithm, summarized in Section 5). We demonstrate that our framework is highly robust to deceptive agents and generalizes to trust evaluation scenarios in many types of networks. Section 6 presents results on the effects of network factors such as homophily and degree, the use of different types of relational features, and robustness to increasing amounts of deception. In Section 7 we discuss related work in the area and conclude in Section 8.

## 2 Informative Networks

Network structure can be intrinsically informative when social forces affect the probability of link formation. Human networks often possess the property of homophily, an increased propensity for like-minded individuals to be connected, colloquially described with the phrase "birds of a feather flock together" [27]. Homophily in trust levels could be categorized as a form of value homophily, the tendency of humans to preferentially connect with people who share the same attitudes and beliefs. Along with value homophily, status homophily, preferential linkages created on the basis of attributes such as age, gender, or ethnicity, is commonly observed in human social networks [20]. Network research has shown that the homophily principle creates strong interpersonal network ties in a wide variety of contexts (e.g., neighborhoods, communities, schools) and affects the choice of informal trusted contacts selected for advice and social support [42]. Clearly, since it is often beneficial for deceptive agents to maintain connections with a network of "dupes", heterophily in trust levels (connections to dissimilar agents) will also exist in trust networks.

A second factor affecting the probability of link maintenance is the agents' satisfaction with past transactions. In most situations, it is reasonable to assume that agents will preferentially maintain connections with trustworthy agents since those relationships are likely to result in direct benefits [35]. Additionally, agents will form and maintain relationships of convenience driven by factors such as proximity, interaction costs, and supply/demand constraints that are not simply explained by either link prediction model [14]. Regardless of these additional factors, we believe that the network structure remains an informative source of information when either value homophily or transactional satisfaction affect link formation.

An underlying assumption of traditional classification methods is that the instances are independent of each other. On the other hand, networks of agents contain instances of multiple types that are related to each other through different types of links. To classify, or label the node in the network, three classification methodologies have been studied over the last decade. Traditional classifiers, often referred to as the content-only classifier, ignore the network and utilize attribute dependencies to predict the label of unknown instances. Relational classifiers improve classification performance

by taking advantage of dependencies of both attribute and labels between related labeled instances [24]. Finally, collective classification aims to simultaneously classify related instances to determine the label of the test node [31, 25]. Studies in other domains have shown that collective classification can increase classification accuracies over non-collective methods when instances are interrelated [30, 41, 22, 47].

## 3 Problem Formulation

Consider the following scenario. An individual agent in a large, open multi-agent system would like to create the largest possible coalition of trustworthy agents for a joint venture. The agent can access the following information:

1. observable features correlated with the agents' trustworthiness;
2. the existence of links connecting agents that have a history of past transactions (but without weights or valences denoting the outcome of the transactions);
3. a set of labels containing information about the trustworthiness of select members of the community.

Note that each link is meant to serve as summary of past transactions rather than representing the outcome of a single transaction. The agent forming the coalition cannot take any probing actions before making its decision. It is assumed that deceptive agents in the system attempt to foil the trust evaluation by two mechanisms:

1. emitting deceptive features;
2. modifying their labels to appear more trustworthy.

For verisimilitude, the network is assumed to follow a power law degree distribution like many human networks, and link formation is driven by a combination of value homophily, transactional satisfaction, and randomness. As a result, there exists a society of $N$ agents connected by graph $G$. In this graph the set of nodes, $V = \{V_1, \ldots, V_n\}$, represents the agents; agents are connected by directed links based on the underlying interactions between the agents. The agents' behavior during interactions is modulated by their own internal value system or *trustworthiness*. The true level of an agent's trustworthiness is hidden from the other agents and can assume a label from the set $L = \{L_1, \ldots, L_n\}$.

Each agent $i$, has two types of attributes: 1) a static feature vector, $S_i = \{s_1, \ldots, s_m\}$, of length $m$; and 2) a dynamic or relational feature vector, $R_i = \{r_1, \ldots, r_n\}$, of length $n$. The static feature vector is observable to all the agents and is related to the agent's trustworthiness; example features could include properties such as "returns library books", "answers email promptly", or "reciprocates invitations". Dynamic, relational features, are calculated through aggregating any known labels of connected agents. The set of agents, $N$, is further divided into two sets of agents: $X$, the agents for whom we know labels (acquaintances or people known by reputation), and $Y$, the agents whose label or trust level need to be determined (strangers). Our task is to determine the labels of the unknown agents, $Y$, from the label set $L$, based on their two types of attributes. The ultimate goal of the observing agent is to recognize the trustworthiness of other agents in the graph and to form a coalition consisting of the most trustworthy set of agents.

## 4 Agent Network Generation

To evaluate the performance of collective classification on identifying agents' trustworthiness in a variety of networks, we simulate the evolution of agent networks formed by the combined forces of value homophily and transactional satisfaction. Since social communities often form a scale-free network, whose degree distribution follows a power law [1], we model our agent networks in the same fashion.

Following the Sen et al. [38] network data generation method, we control the link density of the network using a parameter, *ld*, and value homophily between agents using a parameter, *dh*. The effects of value homophily is simulated as follows:

1. At each step, a link is either added between two existing nodes or a new node is created based on the link density parameter (*ld*). In general, linking existing nodes results in a higher average degree than adding a new node.
2. To add a link, we first randomly select a node as the source node, $A$, and a sink node, $B$, based on the homophily value (*dh*), which governs the propensity of nodes with similar trustworthiness values to link. Node $B$ is selected among all the candidate nodes in the correct class, based on the degree of the node. Nodes with higher degree have a higher chance to be selected.

Transactional satisfaction also governs the process of link formation. Once the link generation process starts, we add a directed link from node $A$ to node $B$ by default, under the assumption that the first selected agent initiated the transaction. The transactional trustworthiness of the second node governs whether a reciprocal link is formed. Here, we use an evaluation function $F_x(p,t)$ to map an observed performance value $p$ in a particular task $t$ to a binary evaluation of performance (positive or negative). We assume that all agents use the same evaluation function for all tasks, which is:

$$F_x(p,t) = \begin{cases} 1: & p \geq 0.5 \\ -1: & p < 0.5 \end{cases}$$

To generate a new node, we first select a trustworthiness level based on a uniform class distribution and assign that class label to the node. Then we add links between the new node and one of the existing nodes as we described above. Inspired by the model proposed by Burnett et al. [5], the trustworthiness label (Table 1(b)) governs the mean and standard deviation parameters of a Gaussian distribution from which simulated performance values are drawn. The algorithm for simulating the evolution of the agent network is outlined in Table 1(a).

After generating the network, we assign observable static attributes to each agent by drawing from a set of binomial distributions based on its trustworthiness. Attributes are represented as a binary feature vector, which indicates the existence or absence of a given feature. These features are meant to represent observable properties that result from the consistent practice of an agent's trust value system. Observable attributes for each class are generated using a set of binomial distributions. Attributes are represented by a binary feature vector, length 10, but the maximum number of attributes that can be true is capped at 5. Random noise is introduced to the attribute generation process using the *attrNoise* parameter. Specifically, with a probability of *attrNoise*, each binary feature is independently assumed to be corrupted, in which case it is set randomly to either 0 or 1 with equal probability. The *attrNoise* parameter can be used to model the

**Table 1.**

(a) Agent Network Generator

**Agent Network Generator (*numNodes*, *ld*, *numLabels*, *attrNoise*, *dh*)**
$i = 0$
G = NULL
while $i <$ *numNodes* do
  sample $r$ from uniform distribution $U(0, 1)$
  if $r \leq ld$ then
    connectNode(G,*numLabels*,*dh*)
  else
    addNodes(G,*numLabels*,*dh*)
    $i = i + 1$
  end if
end while
for $i$ = 1 to *numNodes* do
  Attributes = genAttr($v$, *Attributes*, *label*, *attrNoise*)
  where $v$ is $i$th node in G
end for
return G

(b) Agent Task Performance Profile

| Trust Level | Mean | StDev |
|---|---|---|
| *L1* | 0.9 | 0.05 |
| *L2* | 0.6 | 0.1 |
| *L3* | 0.4 | 0.1 |
| *L4* | 0.2 | 0.05 |

level of deceptiveness of agents in attempting to hide observable attributes that provide clues about their trustworthiness.

## 5 Iterative Classification Algorithm

In this agent network scenario, collective classification refers to the combined classification of a set of interlinked nodes using three types of correlations [39]: 1) correlations between the label of node $V$ and its observed attributes; 2) correlations between the label of node $V$ and the observed attributes (including observed labels of nodes in its neighborhood); and 3) the correlations between the label of node $V$ and the unobserved labels of agents in its neighborhood. For our experiments, we use the iterative classification algorithm [30], an approximate inference algorithm that has shown promise at hyperlink document classification tasks.

Iterative classification was first proposed by [30] and has since been extended by [26]. In ICA, the training model is built using both static and relational attributes of the observed nodes. Since the class labels of the training nodes are known, the value of the dynamic attributes can be calculated using aggregation operators such as *count*, *proportion*, or *mode*. Aggregation operators are different ways of representing the same information (the labels of the connected nodes), but alternate representations have been shown to impact classification accuracy, based on the application domain [39].

The training model is applied to the test nodes whose class labels are unknown; in our problem, these are the stranger agents, for whom no reputation information exists. Initially, because some of class labels of the related nodes are unknown, the values of their relational attributes are also unknown. This problem can be solved by bootstrapping the classification process. At the beginning, the prediction of the class labels for all test nodes is obtained using content features only. Predictions made with high probability are accepted as valid and are accepted into data as known class labels. After certain percentage of classification with highest probability are accepted, the classifier recalculates the relational attributes using the newly accepted labels and reclassifies the

labels. In each iteration, a greater percentage of classifications are accepted and new dynamic attributes are filled in. It is worth noting that the prediction is both recalculated and reevaluated in each iteration; hence the prediction about a given node might change over the process of iteration. Therefore, the label of a node accepted in one iteration might be discarded in the next iteration if the probability associated with the prediction is no longer in the top percentage of acceptance predictions. ICA has the potential to subsequently improve classification accuracy on related data after iterations. However, it should be carefully applied since the incorrect relational features in one iteration may diminish the classification accuracy. Table 2(a) shows the pseudo-code for ICA.

Experiments have shown improvement in classification accuracy by making certain modifications to basic ICA. For instance, [26] proposes a strategy where only a subset of the unobserved variables are utilized as inputs for feature construction. More specifically, in each iteration, they choose the top $K$ most confident predicted labels and use only those unobserved variables in the following iterations predictions, thus ignoring the less confident predicted labels. In each subsequent iteration they increase the value of $K$ so that in the last iteration all nodes are used for prediction.

In this paper, we explore the use of a reputation-based aggregation operator. For a rational agent, its reputation in a trust system is often calculated based on evidence consisting of its observable positive and negative experiences [43]. This evidence can be collected by an agent locally or via a reputation agency. We define the agent's reputation as the average judgment by its observable direct interactions. We assume that the agent will receive a positive evaluation only if its interactors's trust level is equal or lower than itself's. The agent's reputation is therefore the frequency of positive opinions. Suppose $r_x^{N_x}$ is the number of positive evaluation agent $x$ received from its observable interactors $N_x$, and $s_x^{N_x}$ is the number of negative evaluations. We compute reputation based on $r_x^{N_x}$ and $s_x^{N_x}$ as

$$R_x = \frac{r_x^{N_x}}{r_x^{N_x} + s_x^{N_x}}. \tag{1}$$

Note that $R_x$ is a single scalar value, unlike typical aggregation operators such as *count* or *mode*.

## 6 Experiments

Our experimental methodology can be summarized as follows. We generate agent networks using the procedure described in Section 4 with the network parameter values specified in Table 2(b). *numNodes* refers to the total number of agents in the network, including both agents whose trust levels are revealed (analogous to the training set) and those for which trust levels are hidden (corresponding to a test set); *dh* denotes the homophily of the network; *numLabels* is the number of discrete trust levels, with 1 corresponding to the most trustworthy agents; *numFeatures* is the dimensionality of the binary feature vector; *attrNoise* controls the probability that a given binary feature is randomized (corresponding to a degree of deception). Unless indicated otherwise, these parameter values are fixed across experiments and plot classification accuracy against the link density of generated networks. For each network instance, we perform three-fold cross-validation (using disjoint subsets of agents with revealed and hidden labels) and report averaged results.

To evaluate the performance of collective classification in defining the trust level of unknown agents, we adopt the ICA algorithm [26] and employ the Logistic Regression Classifier (LRC) as the baseline classifier in all the experiments.

**Table 2.**

(a) ICA

| **Iterative Classification Algorithm** |
|---|
| 1. Build model on fully labeled training set. |
| 2. Apply trained model to test set of $N$ instances. |
| For each iteration $i$ : 1 to $K$ |
| a. Calculate values for dynamic relational attributes |
| b. Use model to predict class labels |
| c. Sort inferences by probability |
| d. Accept $m$ class labels, where $m = N \times (i/K)$ |
| 3. Output final inferences made by model on test set |

(b) Parameter settings

| **Parameter Name** | **Value** |
|---|---|
| *numNodes* | 500 |
| *dh* | 0.8 |
| *numLabels* | 4 |
| *numFeatures* | 10 |
| *attrNoise* | 0.2 |

We perform a series of experiments to investigate several key issues in collective classification for trust evaluation. First, we compare collective classification against a baseline classifier, both in terms of overall accuracy and on inter-class misclassification. We then explore how the benefits of collective classification depend on network characteristics, such as link density and homophily. We also evaluate the impact of a variety of aggregate operators that represent the relations between trust levels of connected agents and finally examine the robustness of collective classification to two forms of deception in agent networks.

### 6.1 Comparisons against baseline classifier

Figure 1 compares the classification accuracy of ICA against the baseline classifier (logistic regression) for default agent network parameter settings. The feature vector for the baseline algorithm is simply the list of observable binary features, while that of ICA is augmented by the agent's relational attributes expressed using the *count* operator. The latter is a histogram over trust levels of the agents connected to the given agent, computed in both directions (i.e., an additional 8-dimensional feature). As can be seen from the graph, ICA improves over the baseline in a small number of iterations and converges rapidly. Based on this, we use the same value of $K = 10$ for the number of ICA iterations. More importantly, we observe that ICA dramatically improves the classification accuracy from a baseline of 73% to 95%, showing that collective classification is able to exploit significant information about agent trust levels encoded in the network, beyond that expressed in the observable features alone.

Tables 3 presents the confusion matrices for the baseline (LRC) and collective classification (ICA) approaches. We can make several observations about the misclassifications. First, collective classification virtually eliminates the possibility of misclassifying an agent as very untrustworthy (L4). Second, the classification accuracy for L1–3 agents improves dramatically. Finally, although the classification accuracy of L4 agents remains unchanged, we see that ICA is much less likely to misclassify L4 agents as trustworthy (L1).

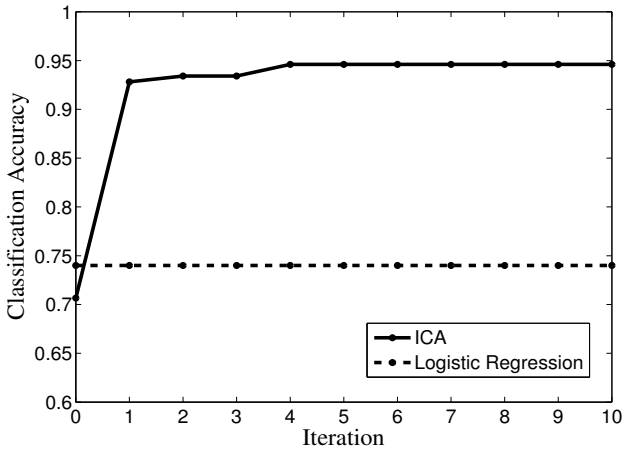



**Fig. 1.** Collective classification (ICA) clearly outperforms the baseline (LRC) and converges in a few iterations. (*ld*=0.4, *dh*=0.8,*attrNoise*=0.2).

**Table 3.** Confusion matrix for baseline (on left) and collective classification (on right) with parameter setting *ld*=0.4,*dh*=0.8,*attrNoise*=0.2

| | ***L1*** | ***L2*** | ***L3*** | ***L4*** |
|---|---|---|---|---|
| ***L1*** | 80.0 | 14.3 | 5.7 | 0 |
| ***L2*** | 16.3 | 60.5 | 20.9 | 2.3 |
| ***L3*** | 2.6 | 5.1 | 74.4 | 17.9 |
| ***L4*** | 6.0 | 0 | 6.0 | 88.0 |

| | ***L1*** | ***L2*** | ***L3*** | ***L4*** |
|---|---|---|---|---|
| ***L1*** | 97.1 | 0 | 2.9 | 0 |
| ***L2*** | 9.3 | 90.7 | 0 | 0 |
| ***L3*** | 0 | 2.6 | 97.4 | 0 |
| ***L4*** | 2.0 | 2.0 | 8.0 | 88.0 |

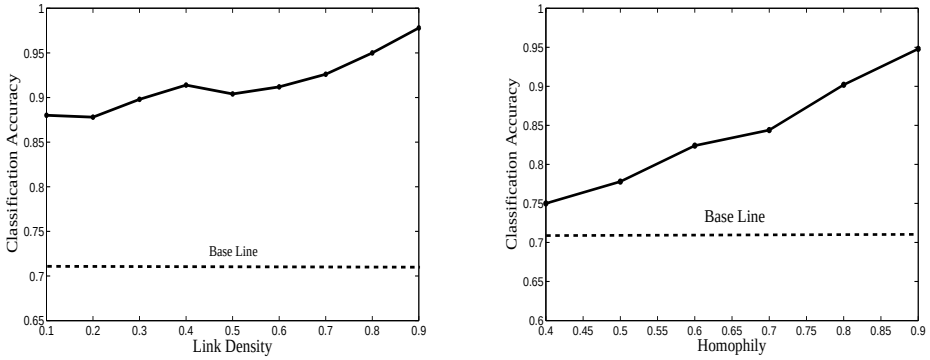
### 6.2 Link density and Homophily

In order to observe the impact of network's link density parameter on collective classification, we generate networks with *ld* changing from $0.1$ to $0.9$ with step size 0.1, and freezing *attrNoise* and *dh* at 0.2 and 0.8, respectively. Figure 2(b) shows how ICA classification accuracy varies with link density. The results show that ICA continues to outperform the baseline and that classification accuracy improves with increased link density. These results are consistent with our expectation that where reliable dependencies exist between instances, increasing the degree of links enables collective classification to more reliably extract relational information from the noisy data, thus improving classification accuracy.

We would also expect collective classification to perform better in networks that exhibit higher levels of homophily. Figure 2(a) shows how classification accuracy varies with different values of homophily (*dh* ranging from $0.4$ to $0.9$ with step size of $0.1$). The results match our predictions: when homophily is low ($dh = 0.4$), the relational information only improves classification results slightly; but as we increase homophily, collective classification accuracy climbs steadily.

### 6.3 Aggregation Operators

Aggregation operators summarize the visible trust levels in a given agent's network neighborhood. In this set of experiments we explore the degree to which classification accuracy is affected by the choice of operator. We consider the following operators, each detailed below: *count*, *proportion*, *mode* and *reputation*.




(a) Changing link density with parameter setting *dh*=0.8,*attrNoise*=0.2

(b) Changing homophily with parameter setting *ld*=0.4,*attrNoise*=0.2

**Fig. 2.** The effect of changing link density (a) and homophily (b) on collective classification accuracy

As described earlier, *count* aggregates trust level labels of neighbors into a histogram of raw counts. *Proportion* is a normalized version of the *count* histogram. *Mode* retains only the most popular trust level, ranging from 1–4. Finally, *reputation* (as given in Equation 1) summarizes the agent's neighborhood in a single scalar quantity and can also be employed as an aggregation operator.

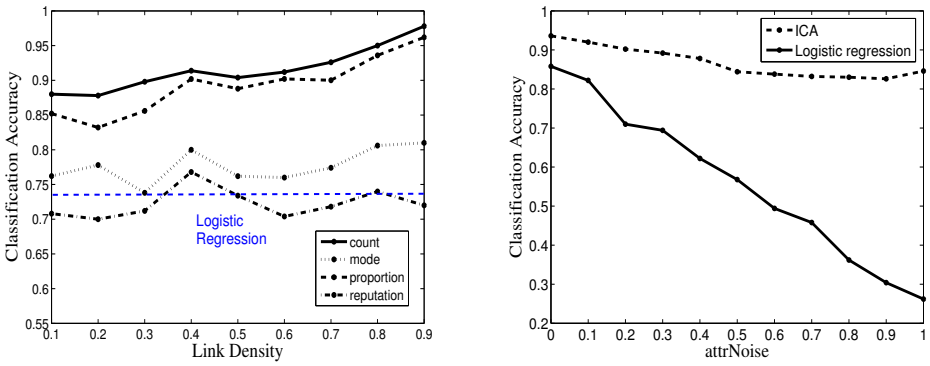
Figure 3(a) compares the classification accuracy of collective classification using the different aggregation operators against the LRC baseline. From the results, we make the following observations. First, compressing the relational information as a single scalar-valued *reputation* does not improve accuracy over the baseline. The *mode* operator is a little better, slightly but consistently outperforming the baseline. However, losing the richness of the visible trust levels (retaining only the most popular) is clearly inferior to the complete histogram of *proportion* or *count*. In fact, the unnormalized counts give the best results, and are therefore used as the default aggregation operator.

### 6.4 Robustness to Deception

So far, we have enforced a completely positive correlation between the agent's feature and its class label (trust level). However, in reality, cases may exist when certain untrustworthy individuals misrepresent themselves by modifying their observable information. In order to evaluate the performance of our model when this assumption is relaxed, we conduct two series of experiments. In the first experiment, we deliberately assign an increasing percentage of the deceptive nodes into the training dataset.

Here, the deceptive agent modifies its class label to appear more trustworthy (i.e., changing from L4 towards L1). Consequently, we select deceptive agents from classes *L2*, *L3*, and *L4*. We run 20 trials for each deception experiment with variable link density. Figure 4 shows the averaged results.

Collective classification (ICA) shows great robustness in this test (see Figure 4 and Table 4 . In a network with a modest amount of homophily, even when a large fraction of the population is deceptive (25% deceivers) ICA can continue to provide reliable results. It is important to note that employing collective classification on even a highly deceptive




(a) *dh*=0.8,*attrNoise*=0.2 (b) *dh*=0.8, *ld*=0.4

**Fig. 3.** Classification accuracy using (a) different aggregation operators; and (b) different noise values on synthetic trust dataset (*attrNoise*)

network is better than ignoring network information (ICA outperforms baseline of 75% in all conditions).

**Table 4.** Even with a high fraction of deceivers, using relations improves over the LRC baseline (75%). (*ld*=0.4,*dh*=0.8,*attrNoise*=0.2)

| Deceivers (%) | 1 | 5 | 10 | 15 | 20 | 25 |
|---|---|---|---|---|---|---|
| Accuracy (%) | 91.2 | 90.3 | 89.5 | 88.7 | 87.6 | 84.2 |

We also seek to explore the robustness of collective classification to a second form of deception: where the agent corrupts its observable features, generating a noisy observation vector. In our network generation model, the *attrNoise* parameter precisely captures the effect: each binary feature is randomized i.i.d. with a probability of *attrNoise*. As in earlier experiments, we compare collective classification (ICA) against the baseline (LRC), as shown in Figure 3(b). We make several observations. First, unlike in previous experiments, we confirm that the baseline accuracy decreases steadily as *attrNoise* rises, reaching chance level (25%) when $attrNoise = 1$. This is because an agent's observable features become an increasingly unreliable predictor of its trustworthiness. Second, by contrast we see that ICA's accuracy degrades surprisingly little, even when observable features become completely non-informative. This is because collective classification is still able to rely on network relations to predict an agent's trustworthiness based solely upon that of other agents in the neighborhood. Clearly, this can happen only when the network exhibits sufficient homophily and density.

## 7 Related Work

Trust evaluation has been applied to many diverse domains including peer-to-peer networks [18, 46], online social networks [48, 28, 34], e-business [29, 32] and mobile adhoc networks [4]. Identifying non trustworthy agents in multi-agent systems and coping

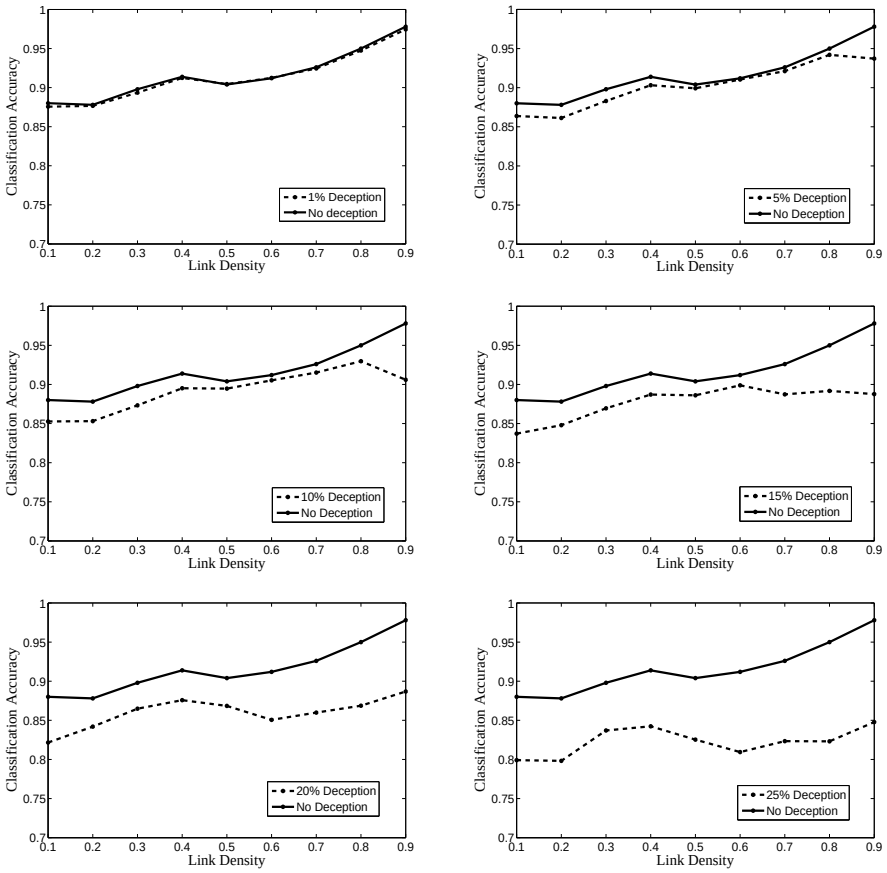



**Fig. 4.** Deception experiment using collective classification with the number of deceivers changing from 1% to 25%.

with the problem of cheating is important especially for the web and in electronic marketplaces. [7, 19] and [40] have proposed techniques to cope with cheaters and sneakers respectively. In our work, we are not only interested in identifying untrustworthy agents, but also finding highly trustworthy agents. Our approach uses local network information to perform a trust evaluation of other agents. In huge networks such as the Semantic Web, this local approach is also favored as the agents do not have access to all other agents. [48] offers some local metrics for trust and reputation in the Semantic Web domain.

Other authors have examined the relationship between trust and homophily in human social networks. Prisel and Anderson [33] observe that perceived homophily is positively related to feelings of safety and is negatively related to the level of uncertainty in groups. Evans and Wensley [9, 10] showed a direct link between homophily and trust; higher levels of status and value homophily increase the level of trust. They

also note that homophily results in increased knowledge/information sharing activities across the group which are often a precursor to trust. However, status homophily has also been found to be negatively related to trust. In [21], the authors found no significant effect of status homophily on benevolence-based trust; age similarity was found to have a negative effect on competence-based trust. Overall, we believe that the link between trust and homophily is an interesting problem worthy of further study.

Our proposed trust evaluation approach identifies the correct label for all of the unlabeled agents in the network; this is the fundamental task of within-network classification techniques [8, 23]. Previous authors have looked at the problem of classifying nodes in social networks (e.g., [17, 16]). In these approaches, both network structure information and node class labels are combined to provide new features to improve classification [15]. Much of the previous work on using machine learning to identify the reputation or trust level of agents in a multi-agent system has used more traditional Bayesian methods (e.g., [12, 3]) and ignored the valuable information in the network structure information. We refer to the surveys of Macskassy et al. [23] and [2] for within-network classification techniques that have been used in social networks. Although, within-network classification has been used in fraud detection applications, such as call networks [11, 6], to detect the fraudulent or legitimate entities in the network, it has not been applied to problems of trust and reputation before. We believe that fraud-detection is another potential application for our trust evaluation approach. Our work is novel also in its detailed examination of the effects of agent deception on the classification performance of a collective classifier.

## 8 Conclusion

In this paper, we have demonstrated that when homophily in trustworthiness is a driving factor in the evolution of an agent network, collective classification is an effective mechanism for leveraging the informative powers of the network, even in the presence of other link generation forces such as transactional satisfaction. Although other types of supervised classifiers [44] and relational models of trust [37] have been explored, they do not propagate information across multiple instances to perform trust evaluation. Preserving the distribution of labels through more expressive aggregation operators such as count and proportion is shown to be more effective than the use of the single reputation feature that encodes the value differential between the trustworthiness of a node and its neighbors. In future work, we are particularly interested in applying this framework toward two types of problems: 1) using trustworthiness levels to perform link prediction in agent networks; 2) learning multi-dimensional models of trust from performance data.

**Acknowledgments** This research was supported in part by NSF award IIS-0845159.

## References

1. A. Barabasi and E. Bonabeau. Scale-free networks. *Scientific American*, pages 60–69, May 2003.
2. S. Bhagat, G. Cormode, and S. Muthukrishnan. Node classification in social networks. *Arxiv preprint arXiv:1101.3291*, 2011.

3. J. Braams. Filtering out unfair ratings in Bayesian reputation systems. *The Icfain Journal of Management Research*, 4(2):48–64, 2005.
4. S. Buchegger and J. Le Boudec. A robust reputation system for mobile ad-hoc networks. *Proceedings of P2PEcon, June*, 2004.
5. C. Burnett, T. J. Norman, and K. Sycara. Bootstrapping trust evaluations through stereotypes. In *9th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 241–248, May 2010.
6. C. Cortes, D. Pregibon, and C. Volinsky. Communities of interest. *Intelligent Data Analysis*, 6(3):211–219, 2002.
7. C. Dellarocas. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49(10):1407–1424, 2003.
8. C. Desrosiers and G. Karypis. Within-network classification using local structure similarity. *Machine Learning and Knowledge Discovery in Databases*, pages 260–275, 2009.
9. M. Evans and A. Wensley. The influence of network structure on trust: Addressing the interconnectedness of network principles and trust in communities of practice. In *The 9th European Conference on Knowledge Management: ECKM 2008*, page 183. Academic Conferences Limited, 2008.
10. M. Evans and A. Wensley. Predicting the influence of network structure on trust in knowledge communities: Addressing the interconnectedness of four network principles and trust. *Electronic Journal of Knowledge Management*, 7(1):41–54, 2009.
11. T. Fawcett and F. Provost. Adaptive fraud detection. *Data mining and knowledge discovery*, 1(3):291–316, 1997.
12. R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Social network classification incorporating link type values. In *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on*, pages 19–24. IEEE, 2009.
13. T. Huynh, N. Jennings, and N. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.
14. M. Jackson. Social and economic networks, 2008.
15. T. Kajdanowicz, P. Kazienko, and P. Doskocz. Label-dependent feature extraction in social networks for node classification. *Social Informatics*, pages 89–102, 2010.
16. T. Kajdanowicz, P. Kazienko, and P. Doskocz. A method of label-dependent feature extraction in social networks. *Computational Collective Intelligence. Technologies and Applications*, pages 11–21, 2010.
17. T. Kajdanowicz, P. Kazienko, P. Doskocz, and K. Litwin. An sssessment of node classification accuracy in social networks using label-dependent feature extraction. *Knowledge Management, Information Systems, E-Learning, and Sustainability Research*, pages 125–130, 2010.
18. S. Kamvar, M. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*, pages 640–651. ACM, 2003.
19. R. Kerr and R. Cohen. Smart cheaters do prosper: Defeating trust and reputation systems. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 993–1000. International Foundation for Autonomous Agents and Multiagent Systems, 2009.
20. P. Lazarsfeld and R. Merton. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18:18–66, 1954.
21. D. Levin, R. Cross, and L. Abrams. Why should I trust you? Predictors of interpersonal trust in a knowledge transfer context. In *Academy of Management Meeting, Denver, CO*, 2002.
22. Q. Lu and L. Getoor. Link-based classification. In *In proceedings of 20th International Conference on Machine Learning*, pages 496–503. Association for Computing Machinery, August 2003.

23. S. Macskassy and F. Provost. A brief survey of machine learning methods for classification in networked data and an application to suspicion scoring. In *Proceedings of the 2006 conference on Statistical network analysis*, pages 172–175. Springer-Verlag, 2006.
24. S. A. Macskassy and F. Provost. A simple relational classifier. In *Proceedings of the Second Workshop on Multi-Relational Data Mining (MRDM-2003) at KDD-2003*, pages 64–76, August 2003.
25. S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. In *Journal of Machine Learning Research*, pages 935–983. Association for Computing Machinery, January 2007.
26. L. K. Mcdowell, K. M. Gupta, and D. W. Aha. Cautious inference in collective classification. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 596–601. AAAI Press, July 2007.
27. M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27:415–444, 2001.
28. L. Mui. *Computational models of trust and reputation: Agents, evolutionary games, and social networks*. PhD thesis, Massachusetts Institute of Technology, 2002.
29. L. Mui, M. Mohtashemi, and A. Halberstadt. A computational model of trust and reputation. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*, pages 2431–2439. IEEE, 2002.
30. J. Neville and D. Jensen. Iterative classification in relational data. In *Proceedings of the AAAI 2000 Workshop Learning Statistical Models from Relational Data*, pages 42–49, July 2000.
31. J. Neville and D. Jensen. Collective classification with relational dependency networks. In *Proceedings of KDD-2003 Workshop on Multi-Relational Data Mining (MRDM-2003)*, pages 77–91. AAAI Press, August 2003.
32. J. ODonovan, B. Smyth, V. Evrim, and D. McLeod. Extracting and visualizing trust relationships from online auction feedback comments. In *Proc. IJCAI'07*, pages 2826–2831, 2007.
33. M. Prisbell and J. Andersen. The importance of perceived homophily, level of uncertainty, feeling good, safety, and self-disclosure in interpersonal relationships. *Communication Quarterly*, 28(3):22–33, 1980.
34. J. Pujol, R. Sanguesa, and J. Delgado. Extracting reputation in multi agent systems by means of social network topology. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pages 467–474. ACM, 2002.
35. S. Ramchurn, D. Huynh, and N. Jennings. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19(01):1–25, 2004.
36. P. Resnick and R. Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *Advances in Applied Microeconomics: A Research Annual*, 11:127–157, 2002.
37. A. Rettinger, M. Nickles, and V. Tresp. A statistical relational model for trust learning. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, pages 763–770. International Foundation for Autonomous Agents and Multiagent Systems, 2008.
38. P. Sen and L. Getoor. *Link-based classification*. Technical Report, CS-TR-4858, University of Maryland, Reading, Massachusetts, 2007.
39. P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-rad. *Collective Classification in Network Data*. AI Magazine, 2008.
40. R. Seymour and G. Peterson. Responding to Sneaky Agents in Multi-agent Domains. In *Proceedings of the Florida AI Research Society Conference (FLAIRS)*, 2009.

41. B. Taskar. Discriminative probabilistic models for relational data. In *In Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence*, pages 485–492. Association for Uncertainty in Artificial Intelligence, August 2002.
42. G. van de Bunt, R. Wittek, and M. de Klepper. The evolution of intra-organizational trust networks. *International sociology*, 20(3):339–369, 2005.
43. Y. Wang and M. P. Singh. Formal trust model for multiagent systems. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1551–1556, January 2007.
44. W. Yuan, D. Guan, Y. Lee, and S. Lee. A trust model with dynamic decision making for ubiquitous environments. In *14th IEEE International Conference on Networks*, volume 1, pages 1–6. IEEE, 2007.
45. G. Zacharia and P. Maes. Trust management through reputation mechanisms. *Applied Artificial Intelligence*, 14(9):881–907, 2000.
46. R. Zhou, K. Hwang, and M. Cai. Gossiptrust for fast reputation aggregation in peer-to-peer networks. *Knowledge and Data Engineering, IEEE Transactions on*, 20(9):1282–1295, 2008.
47. X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning*, 2003.
48. C. Ziegler and G. Lausen. Propagation models for trust and distrust in social networks. *Information Systems Frontiers*, 7(4):337–358, 2005.