Entity and Relationship Labeling in Affiliation Networks

Bin Zhao BZHAO@CS.UMD.EDU
Prithviraj Sen SEN@CS.UMD.EDU
Lise Getoor GETOOR@CS.UMD.EDU
Department of Computer Science, University of Maryland, College Park, MD 20742

Abstract

Many domains are best characterized as an affiliation network describing a set of actors and a set of events interlinked together in a variety of relationships. Relational classification in these domains requires the collective classification of both entities (actors and events) and relationships. We investigate the use of relational Markov networks (RMN) for relational classification in affiliation networks. In this paper, we introduce a novel dataset, Profile in Terror (PIT) knowledge base, that provides a rich source of various affiliation networks. We study two tasks, entity labeling and relationship labeling. We highlight several important issues concerning the effectiveness of relational classification. Our results show that the PIT dataset has a rich source of relational structure and therefore it is a useful dataset for statisical relational network learning community.

1. Introduction

Recently, there has been a surge of interest in research involving social networks. This includes modeling and analyzing networks arising from various domains such as online communities (http://www.myspace.com, http://www.orkut.com) and communication within a community (http://www.cs.cmu.edu/∼enron/). Social networks arising out of the counter-terrorism domain have received special attention as can be seen by the spate of forums calling for research in this domain (Workshops on Link Analysis, Counter-terrorism and Security, 2004, 2005, 2006). One of the problems in this domain is the lack of a real-world dataset that is conducive to experimentation and testing of ideas.

In this paper, we present a multi-relational dataset containing entities related to counter-terrorism that, we hope, will provide researchers with a real-world dataset to work on. This dataset, which we refer to as Profiles in Terror (PIT) (http://profilesinterror.mindswap.org/), contains counter-terrorism intelligence information collected from various publicly available real-world sources such as federal court indictments and news reports. Besides being useful to researchers working on counter-terrorism related topics, the PIT knowledge base contains aspects of interest to many research topics related to statistical relational classification/learning, link analysis and data mining.

In this paper, we also provide a preliminary set of experiments to demonstrate the usefulness of the PIT knowledge base. We performed two sets of experiments each addressing a different relational classification task:

- Entity labeling in affiliation networks.
- Relationship labeling in affiliation networks.

Affiliation networks are a type of social network composed of two types of nodes: a set of *actors* and a set of *events*. The actors are linked to events and each link denotes the participation of that actor in that particular event. For each set of experiments we begin by extracting the relevant affiliation network from the PIT knowledge base and performed relational classification experiments.

Since the PIT dataset is best represented as a relational dataset, we performed experiments classifying samples using both the descriptive attributes available on the input and the link structure provided by the dataset. Recent research in statistical relational learn-

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006. Copyright 2006 by the author(s)/owner(s).

ing and classification for structured output spaces have provided us with a long list of tools for classification of networked data using the correlations present in the attributes and the link structure of the data. Some of these tools are: Relational Markov networks (RMN) (Taskar et al., 2002) and Conditional Random fields (CRFs) (Lafferty et al., 2001), Relational Dependency networks (RDNs) (Neville & Jensen, 2004). All our experiments were performed using RMNs due to their simplicity and efficacy. For comparison, we provide the results returned by a content-only maximum entropy classifier as a baseline to demonstrate that PIT dataset is an interesting dataset for statistical relational network learning community.

The remainder of the paper is organized as follows. Section 2 describes the PIT knowledge base in more detail. Section 3 describes relational Markov networks, the relational classifier we used in our experiments. Section 4 introduces affiliation networks and describes the entity labeling and the relationship labeling problems for affiliation networks. Section 5 describes our experiments with entity labeling on the PIT knowledge base. Section 6 describes our experiments with relationship labeling on the PIT knowledge base. In Section 7, we discuss our findings and conclude with future work in Section 8.

2. Profiles in terror (PIT) Knowledge Base

The PIT knowledge base is an ongoing project that began collecting counter-terrorism related information from June 2004. As part of the project, numerous researchers compiled related information from publicly available sources including online news reports, federal court indictments and various publications on counter-terrorism. The PIT knowledge base is not designed to be a comprehensive store of counter-terrorism intelligence. Much of information in this domain is classified and cannot be included in a publicly available knowledge base such as PIT.

The PIT knowledge base consists of various types of entities. Here is an incomplete list of the different entity types:

- `Terrorist organizations` such as Hamas, Hizballah, Liberation Tigers of Tamil Eelam (LTTE), etc.
- `Terrorists` such as Osama bin Ladin, Ramzi Yousef, etc.
- `Terrorist facilities` such as Darunta Training Camp, Khalden Training Camp, etc.
- `Terrorist events/attacks` such as African embassy bombings of 1998, Madrid Bombings of 2004, etc.

`Terrorist organizations` 95	`Terrorists` 435
`Terrorist facilities` 34	`Terror attacks` 1293

Table 1. Counts of the various entity types in the PIT knowledge base.

Each entity instance is associated with various attribute values. For example, each instance of a `terror attack` has a short, one-line *summary* attribute along with a much more detailed *description* attribute. Besides these, `terror attacks` also have attributes specifying the *date* of attack, *location* of attack and *number of people injured*. The various attacks are also subclassified into the nature of attacks, e.g., *kidnappings*, *bombings* and *arson*.

The dataset also contains various types of relations connecting instances of different entity types. Here is a partial list of the various relation types:

- `memberOf`: instances of `terrorist` can be affiliated with various instances of `terrorist organization`.
- `facilityOwner`: instances of `terrorist facility` are usually run by instances of `terrorist organizations`.
- `facilityMember`: instances of `terrorist` are linked to various instances of `terrorist facilities` if the `terrorist` instance attended/spent some time at the facility.
- `claimResponsibility`: instances of `terrorist organization` are linked to the instances of `terror attacks` they claim responsibility for.
- `participate`: instances of `terrorist` may participate in instances of `terror attacks`.

Table 1 and Table 2 show the latest counts of instances of the various entity types and relation types respectively.

The website http://profilesinterror.mindswap.org/ provides a platform for studying advanced and new methodologies for predictive modeling, terrorist (social) network analysis, and visualization of terrorists activities and relationships using Semantic Web technologies. The ontology of PIT is described in OWL

facilityOwner	memberOf	facilityMember	claimResponsibility	participate
7	320	60	817	69

Table 2. Counts of the various relation instances in the PIT knowledge base. See text for definitions of various relation types.

and the knowledge base is stored in an RDF datastore. This allows the data to be machine-readable and one can easily obtain different views of subsets of the dataset.

One of the challenges in the counter-terrorism domain is to identify hidden threats, predict forthcoming events and correctly understand the relations between terrorists based on the already known partial, incomplete or even inaccurate information. Analysts want to know if there are any patterns by which terrorist events take place, how exactly two terrorists are related, if two terror events are plot by the same terrorist organizations or to predict if two seemingly unconnected terrorists are related. All these can be translated into machine learning tasks such as entity labeling, relationship labeling and relationship prediction.

In this paper, we report results of relational classification experiments on the PIT dataset and we next describe the relational classifier we used for these experiments.

3. Relational Markov Networks

Undirected graphical models or *Markov networks* (Cowell et al., 1999) have been shown to be an effective way to represent diverse classification problems and correlations due to the link structure. Due to the flexibility they offer, all our experiments with the PIT dataset were performed using *Relational Markov networks* (RMNs) (Taskar et al., 2002), an extension of Markov networks to relational domains. Here we review the RMN framework.

Let $\mathbf{V}$ be a set of discrete random variables, and let $\mathbf{v}$ be an assignment of values to the random variables. A Markov network is described by a graph $G = (\mathbf{V}, E)$ and a set of parameters Ψ. Let $C(G)$ denote a set of (not necessarily maximal) cliques in G. For each $c \in C(G)$, let V_c denote the nodes in the clique. Each clique c has a clique potential $\psi_c(V_c)$ which is a nonnegative function on the joint domain of V_c and let $\Psi = \{\psi_c(V_c)\}_{c \in C(G)}$. For classification problems we are often interested in conditional models. Let $\mathbf{X}$ be the set of observed random variables we condition on and let $\mathbf{x}$ denote the observed values of $\mathbf{X}$. Let X_c denote the observed random variables in clique $c \in C(G)$ and let x_c denote the observed values of X_c. Let $\mathbf{Y}$ be the set of target random variables to which we want to assign labels and let $\mathbf{y}$ denote an assignment to $\mathbf{Y}$. Let Y_c denote the set of target random variables in clique $c \in C(G)$ and let y_c denote an assignment to it. A *conditional Markov network* or *conditional random field* is a Markov network (G, Ψ) which defines the distribution $P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{c \in C(G)} \psi_c(x_c, y_c)$ where $Z(\mathbf{x}) = \sum_{\mathbf{y}'} \prod_c \psi_c(x_c, y'_c)$.

Conditional Markov networks, as presented above, are not suited for relational classification tasks since they involve clique specific potentials $\psi_c(V_c)$. RMNs (Taskar et al., 2002) are an extension of the Markov network framework to relational domains where we define the clique potentials in log-space using a small set of feature functions $\log \psi_c(y_c, x_c) = \sum_i w_i f_i(x_c, y_c)$ where f_i is the i^{th} feature function (usually a simple indicator function) and w_i is a parameter which needs to be estimated.

Parameter estimation for RMNs can be performed using gradient-based optimization methods from fully labeled training data (Taskar et al., 2002). Taskar et al. also show that to estimate the gradient one needs to perform inference over the training data. In relational domains, the underlying Markov network is usually large and densely connected making exact inference infeasible. Thus Taskar et al. propose the use of approximate inference methods like loopy belief propagation (Yedidia et al., 2000).

4. Labeling Tasks in Affiliation Networks

In this section we review the concept of an affiliation network from Wasserman and Faust (1994) and define the problems of entity labeling and relationship labeling for affiliation networks.

An affiliation network $(\mathcal{N}, \mathcal{M}, E)$ consists of two types of nodes, a set of *actors* $\mathcal{N} = \{n_1, n_2, \ldots, n_g\}$ and a set of *events* $\mathcal{M} = \{m_1, m_2, \ldots, m_h\}$, and the set of edges E. An actor $n \in \mathcal{N}$ is said to be *affiliated* to an event $m \in \mathcal{M}$ if the actor is a member of the event and this is denoted by introducing a link between the actor n and event m such that $(n, m) \in E$.

4.1. Entity Labeling in Affiliation Networks

One of the tasks we consider for our experiments is *entity labeling* for affiliation networks. Consider classifying actors in a given affiliation network with a given set of labels. We simplify the problem by first constructing a network with only a single type of entity, a network consisting only of actors, from the affiliation network. A simple way to do this is to introduce a link between two actors if they are connected to the same event and having introduced these actor-actor links delete all the event nodes and the links emanating from them from the affiliation network. In other words, we introduce an actor-actor link between actors $n_i, n_j \in \mathcal{N}$ if $\exists m_k \in \mathcal{M}$ such that $(n_i, m_k) \in E$ and $(n_j, m_k) \in E$. Notice that the same approach can be applied to label the event nodes in the affiliation networks. We report experimental results for entity labeling on the PIT dataset in Section 5.

4.2. Relationship Labeling in Affiliation Networks

Our second set of experiments were devoted to labeling the actor-actor links/relations themselves in affiliation networks given a fixed set of labels.

Given an affiliation network we begin by first constructing a network consisting only of the actor nodes using the approached described above. Since we want to label the relations among actors we would like to have a Markov network where the relations are represented by target random variables and edges represent correlations. To obtain such a Markov network we perform a simple inversion of the actor-actor network. Let n_i, n_j and n_k represent actors in the actor graph and let e_{ij} and e_{jk} denote the relations connecting n_i, n_j and n_j, n_k respectively. In the Markov network, we introduce target random variables representing the labels for e_{ij} and e_{jk} and edges connecting every pair of such relations e_{ij} and e_{jk} that have an actor in common (viz. n_j).

The main intuition behind connecting nodes representing relations is to exploit the correlation amongst labels on relations connecting the same actors. Our experiments indicate that relations involving the same actors often have the same labels. Taskar et al. (2004) used a similar approach to classify hyperlinks connecting university webpages.

5. Experimental Results: Entity Labeling

For the first set of experiments we chose the `terror attack` part of the PIT dataset and extracted two different types of affiliation networks to experiment with.

The first affiliation network consisted of `terror attacks` as actors and an event defined for every location. In this affiliation network, denoted by *loc* affiliation network, a `terror attack` is connected to a location node if the attack took place in that location. The second affiliation network also consisted of `terror attacks` representing actors but defined an event to be a pair of location and `terrorist organization`. In the second affiliation network, denoted by *loc+org* affiliation network, a `terror attack` node is connected to an event if and only if the attack is claimed by the organization and took place in that location.

	Flat	RMN loc	RMN loc+org
Avg. Accuracy	87.06	86.93	87.1

Table 3. Average classification accuracy of terror attacks

In the PIT knowledge base, there are a total 1,293 `terror attack` instances each classified into one of six classes denoting the type of attack: *arson* (2.4%), *bombing* (43.5%), *kidnapping* (13.8%), *NBCR* (stands for Nuclear, Biological, Chemical or Radiation attack) (0.6%), *weapon attack* (38.5%) and *other* (1%). We split the extracted networks into three sets each containing around 430 instances to be labeled and performed three-fold cross validation.

Each `terror attack` instance is associated with many descriptive attributes including year of attack, keywords from a description written by a human etc. We used these attribute values as evidence during classification. To facilitate relational classification, we used RMNs to classify the instances using the descriptive attributes as well as the link structure provided by the affiliation networks. As a baseline, we compare the various RMNs against a content-only maximum entropy classifier (flat model) that classifies using only the descriptive attributes.

For each classifier, we assume a "shrinkage" prior and compute the MAP estimate of the parameters. More precisely, we assumed that different parameters are a priori independent and define $p(w_i) = \lambda w_i^2$. We tried a range of regularization constants for each classifier and found that $\lambda = 10$ returned the best results. Taskar et al. (2002) report using a regularization constant of the same magnitude $\lambda \approx 5.5$.

As Table 3 shows, the RMNs and the flat model return almost identical performance and there is not much to choose between them. The reason for this turns out to be the high quality of evidence we considered. Each terror attack is accompanied by a description that was written by a human while entering the terror attack into the knowledge base. This description frequently contains some highly predictive words, e.g., "explosion" or "detonated" in the case of a bombing etc.

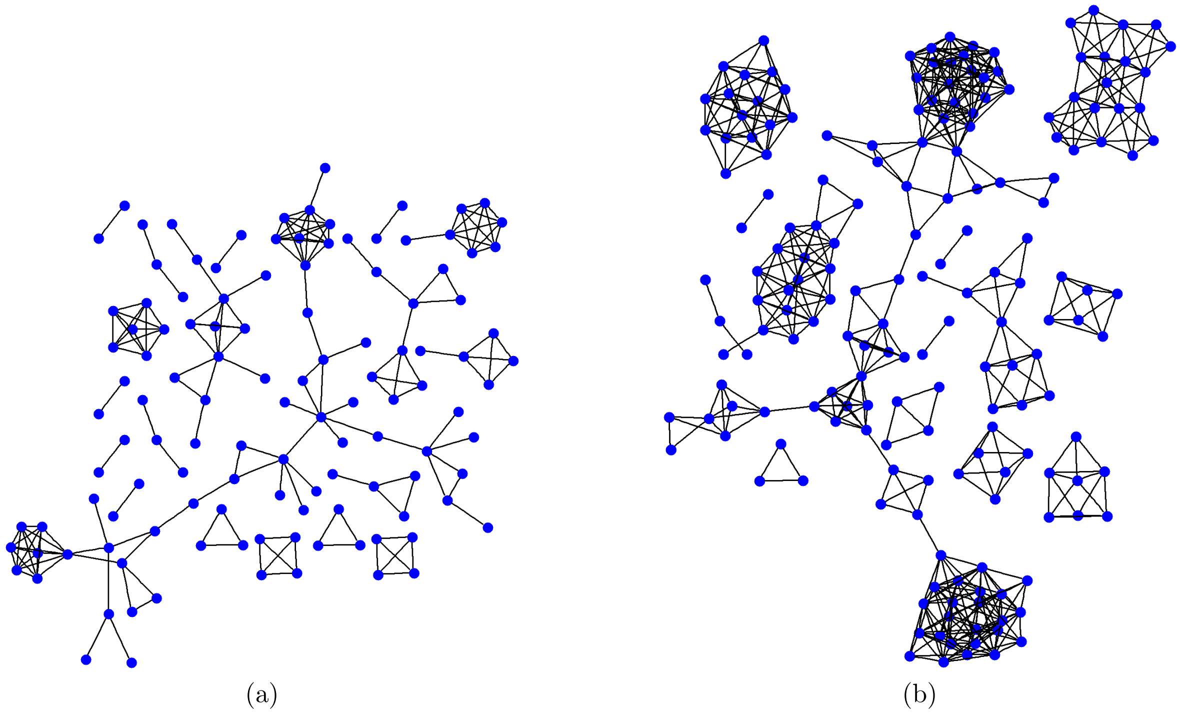

Figure 1. (a) A terrorist graph from the PIT dataset consisting of 181 edges and (b) the corresponding link graph consisting of 580 edges

In future, we aim to exclude this type of human written evidence and use other type of automatically gathered evidence to find out if machine learning classifiers can label effectively.

6. Experimental Results: Relationship Labeling

At the core of terrorist activity is a network of personal connections that allows the terrorist organization to function. Consequently, looking at who knows whom and how they are related to each other is central to understanding the extent of terrorist activities. Intelligence information can show that two terrorists are related in some ways but in what exact way is often unknown. Therefore it is important to understanding the nature of the relation structure amongst the terrorists from the known data and be able to label all the unknown relations.

For the second set of experiments we extracted an affiliation network from the `terrorist` subset of the PIT knowledge base. This affiliation network consists of `terrorist` instances as actors and institutions representing events, where an institution can be one of terrorist organization, family, communication or users of the same terrorist facility. `Terrorists` link to events if they are members of the same institution. We transformed the affiliation network to obtain an actor-actor graph (with 917 edges) and, finally, to a Markov network as described in Section 4.2. Figure 1 shows a subset of the actor-actor graph and its corresponding Markov network. Recall that in this set of experiments we aim to label the actor-actor relationships themselves. More specifically, we would like to see if we can assign each relationship its correct set of labels where the set of labels is:

- *accomplice* (53.1%): An accomplice relation means two people are members of the same terrorist organization.

- *family* (14.8%): A family relation means two people are in the same family (e.g. father-son, husband-wife, uncle-nephew, cousin-cousin).

- *contact* (19.6%): A contact relation means two people have contacted each other (e.g. attend the same meeting, email each other, call each other via phone).

- *congregate* (12.4%): A congregation relation means two people use the same facility (e.g. went to the same training camp).

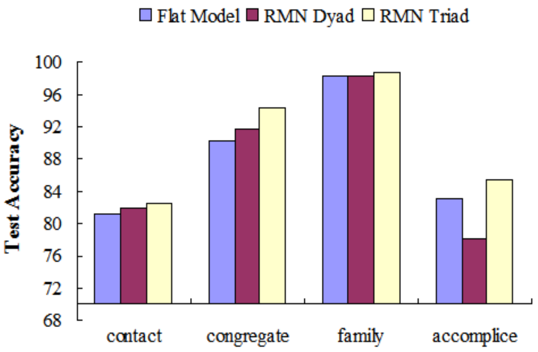

Figure 2. The average classification accuracy for binary terrorist relationship labeling.

Note that the above labels are not disjoint since it is possible that two terrorists be linked both because they belong to the same family and they are members of the same terrorist organization. Thus this problem is an instance of *relational multi-label classification*. We split the `terrorist` dataset into two sets and performed two-fold cross validation.

As part of the evidence, we included with each relationship various types of information belonging to the terrorists involved in the relation such as nationality, text from their biography etc. We report the accuracies of a content-only maxEnt classifier as a baseline (flat).

In this set of experiments we experimented with two different RMNs. In the first one, we included all cliques upto size 2 in the Markov network (*dyad RMN*). The *dyad RMN* usually turned out to consist of too many cliques for RMNs to handle. In particular, approximate inference techniques like loopy belief propagation (Yedidia et al., 2000) are known to provide poor approximations when there are a number of short, closed loops (Yedidia et al., 2005) (a direct consequence of high link density) in the underlying Markov network. Due to the poor quality of inference, the parameter estimation for RMNs often did not converge to desirable values.

In an effort to reduce the number of cliques in the generated Markov networks we also experimented with RMNs where we included all the three-cliques present in Markov network besides node cliques and refrained from including edge cliques (*triad RMN*).

Just as before, for each classifier, we assumed a "shrinkage" prior and compute the MAP estimate of the parameters using a regularization constant of 10.

6.1. Multi-label classification results

We first report the results of the *relational multi-label classification* experiments. A simple way to perform multi-label classification is to learn numerous binary one-against-the-rest classifiers. Thus we learn four different types of classifiers one for each of *accomplice*, *family*, *contact* and *congregate*. The results are shown in Figure 2.

Figure 2 shows that the triad RMN always does better than the flat model. As we remarked earlier, the dyad RMN sometimes (in the case of accomplice) fails to improve upon the results of the flat model due to the excessive link density.

As part of our future work we aim to utilize methods such as Ghamrawi and McCallum (2005) to perform collective multi-label classification.

6.2. Single-label multi-class classification results

As part of our efforts to perform some experiments on multi-class data we obtained a single label dataset by throwing out all the relations with multiple labels. This reduced our dataset from 917 to 884 relations. Thus we obtained a single label multi-class dataset.

The set of bars labeled "All" in Figure 3 shows the results on this dataset for the three classifiers flat, dyad RMN and triad RMN on the multi-class classification problem. Notice that in Figure 3 the dyad RMN performs consistently better than the triad model and the triad RMN consistently improves upon the results of the flat model showing that both dyad and triad cliques can be useful for relationship labeling. One of the reasons for the dyad RMN doing better than the triad RMN could be the fact that there are a lot more cliques in the dyad RMN thus allowing the inference procedure to exploit more correlations that exist in the link structure.

In Figure 3, we also report the results of experiments without certain features. In particular, Taskar et al. (2004) report that relation classification (in their case hyperlinks) may be improved if one includes as part of the evidence the labels on the entities (in their case the webpages) themselves. In Figure 3, the "No Type" results were obtained by not including in the set of features the class labels on the terrorists (leader, terrorist etc.). We confirm that including the labels on the entities as evidence aids the relationship classification. The set of results labeled "No Keywords" were obtained by not utilizing the biographies of the terrorists as evidence which happens to be a substantial part of the feature set and the results show that relational

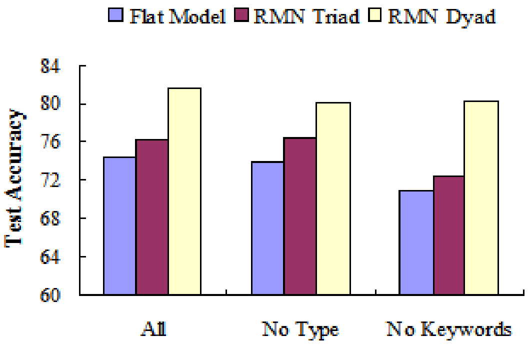

Figure 3. Average classification accuracy of terrorist relation labeling.

methods can do well even when there is a dearth of evidence.

7. Discussion

We were confronted with a number of issues while preprocessing and experimenting with the PIT dataset. Here we discuss each issue in turn with the hope of identifying important avenues for future work.

One of the issues that usually comes up during the pre-processing of relational datasets is how to construct training and test datasets. It is not always the case that the dataset itself provides subsets that are natural splits such as the university splits in WebKB (Craven et al., 1998) where each split forms a disjoint graph. One common approach used to create training and test splits for identically and independently distributed samples is to create randomly sampled stratified subsets of the data so that each subset contains the same distribution of class labels. This approach fails on two counts in the case of relational data:

- Random sampling may cause linked entities to fall into different subsets. The links that go from one subset to another are usually ignored during parameter estimation if both the subsets are not used for learning and this means that we are ignoring some information and not using the data fully.
- The intuition behind creating training and test sets is to make sure that they come from the same distribution. Unfortunately, since random stratified sampling does not look at the links, it may be the case that we construct splits containing an unequal number of links. Figure 4 shows two splits that were created from the terrorist relation PIT dataset. Note that Figure 4 (b) is much denser than Figure 4 (a). Clearly, these two splits do not represent the same distribution.

Another important issue that comes up when dealing with relational datasets is the problem of high link density. Common approximate inference techniques, e. g. loopy belief propagation (Yedidia et al., 2000), face problems when run on datasets containing numerous densely clustered nodes forming short, closed loops (Yedidia et al., 2005). This usually causes the approximate inference approach to return a poor approximation resulting in poor quality inference. Inference for such densely connected datasets is still an interesting open problem.

Our experience with the PIT dataset shows that this dataset is quite different from common relational datasets such as WebKB (Craven et al., 1998) or Cora (McCallum et al., 2000). The PIT dataset contains a larger number of clusters of nodes making it a much more challenging dataset. We hope that such datasets with markedly different properties will help researchers in the field identify new and interesting problems to work on.

8. Conclusion and Future Work

In this paper we introduced a multi-relational dataset, the PIT knowledge base, containing various entities related to counter-terrorism. We described various entity types and relation types present in the data. Our hope is that this dataset will facilitate research in numerous fields including link analysis, statistical relational learning, data mining etc. We also provided a preliminary set of experiments performed on the different social networks, in particular, affiliation networks, that can be extracted from the PIT dataset.

Acknowledgments

We would like to thank Dr. James Hendler for his valuable inputs on this paper. We also would like to thank reviewers for their constructive comments and suggestions. This work was completed as part of work of Maryland Information and Network Dynamics Laboratory Semantic Web Agents Project and with funding from Fujitsu Laboratories of America-College Park, Lockheed Martin Advanced Technology Laboratory, NTT Corp.,Kevric Corp., SAIC, National Science Foundation, National Geospatial-Intelligence Agency, DARPA, US Army Research Laboratory, NIST, and other DoD sources.

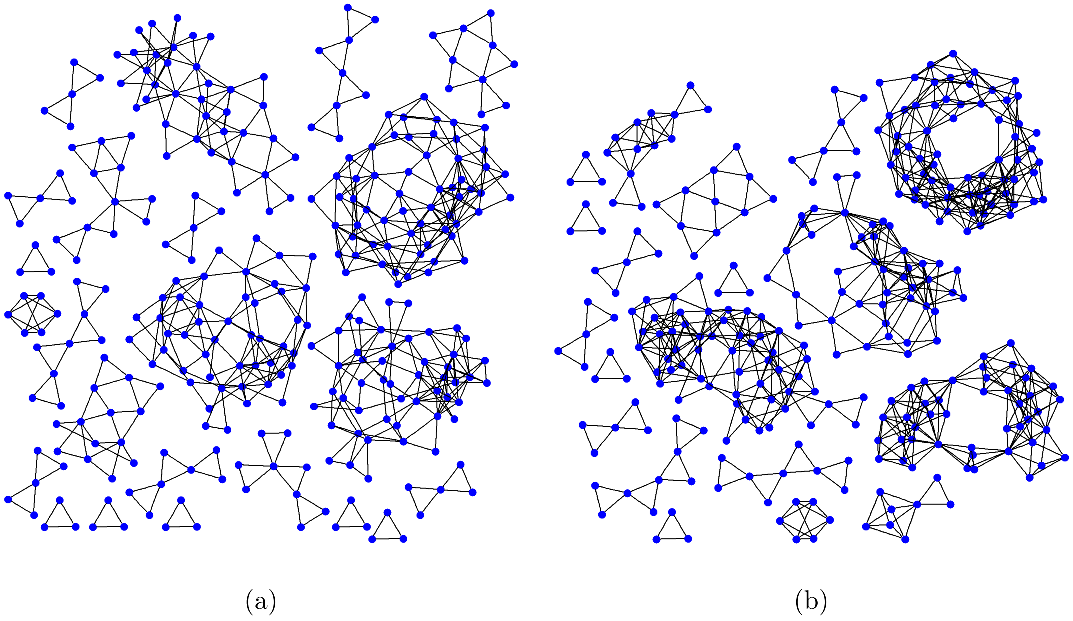

Figure 4. Naively created splits of different densities using stratified sampling. (a) A split comprising of 442 nodes and 225 edges. (b) A split comprising of 442 nodes and 283 edges.

References

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. Springer, New York.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., & Slattery, S. (1998). Learning to extract symbolic knowledge from the world wide web. *Proceedings of the AAAI*.

Ghamrawi, N., & McCallum, A. (2005). Collective multi-label classification. *CIKM*.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. 18th International Conf. on Machine Learning*.

McCallum, A., Nigam, K., Rennie, J., & Seymore, K. (2000). Automating the construction of internet portals with machine learning. *Information Retrieval*.

Neville, J., & Jensen, D. (2004). Dependency networks for relational data. *Proceedings of the IEEE International Conference on Data Mining*.

Taskar, B., Abbeel, P., & Koller, D. (2002). Discriminative probabilistic models for relational data. *18th Conference on Uncertainty in Artificial Intelligence*.

Taskar, B., Wong, M.-F., Abbeel, P., & Koller, D. (2004). Link prediction in relational data. *Neural Information Processing Systems*.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Yedidia, J., Freeman, W., & Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory* (pp. 2282–2312).

Yedidia, J., Freeman, W. T., & Weiss, Y. (2000). Generalized belief propagation. *Neural Information Processing Systems* (pp. 689–695).