Revisiting Role Discovery in Networks: From Node to Edge Roles

Nesreen K. Ahmed
Intel Labs
nesreen.k.ahmed@intel.com

Ryan A. Rossi
Palo Alto Research Center
rrossi@parc.com

Theodore L. Willke
Intel Labs
ted.willke@intel.com

Rong Zhou
Palo Alto Research Center
rzhou@parc.com

ABSTRACT

Previous work in network analysis has focused on modeling the mixed-memberships of node roles in the graph, but not the roles of edges. We introduce the *edge role discovery problem* and present a generalizable framework for learning and extracting edge roles from arbitrary graphs automatically. Furthermore, while existing node-centric role models have mainly focused on simple degree and egonet features, this work also explores graphlet features for role discovery. In addition, we also develop an approach for automatically learning and extracting important and useful edge features from an arbitrary graph. The experimental results demonstrate the utility of edge roles for network analysis tasks on a variety of graphs from various problem domains.

1. INTRODUCTION

In the traditional graph-based sense, roles represent node-level connectivity patterns such as star-center, star-edge nodes, near-cliques or nodes that act as bridges to different regions of the graph. Intuitively, two nodes belong to the same role if they are "similar" in the sense of graph structure. Our proposed research will broaden the framework for defining, discovering and learning network roles, by drastically increasing the degree of usefulness of the information embedded within rich graphs.

Recently, role discovery has become increasingly important for a variety of application and problem domains [8, 17, 5, 4, 30, 24, 37] including descriptive network modeling [31], classification [16], anomaly detection [31], and exploratory analysis (See [30] for other applications). Despite the (wide variety of) practical applications and importance of role discovery, existing work has only focused on discovering node roles (*e.g.*, see[4, 6, 11, 27]). We posit that discovering the roles of edges may be fundamentally more important and able to capture, represent, and summarize the key behavioral roles in the network better than existing methods that have been limited to learning only the roles of nodes in the graph. For instance, a person with malicious intent may appear normal by maintaining the vast majority of relationships and communications with individuals that play normal roles in society. In this situation, techniques that reveal the role semantics of nodes would have difficulty detecting such malicious behavior since most edges are normal. However, modeling the roles (functional semantics, intent) of individual edges (relationships, communications) in the rich graph would improve our ability to identify, detect, and predict this type of malicious activity since we are modeling it directly. Nevertheless, existing work also have many other limitations, which significantly reduces the practical utility of such methods in real-world networks. One such example is that the existing work has been limited to mainly simple degree and egonet features [16, 31], see [30] for other possibilities. Instead, we leverage higher-order network motifs (induced subgraphs) of size $k \in \{3, 4, \ldots\}$ computed from [1, 2] and other graph parameters such as the largest clique in a node (or edge) neighborhood, triangle core number, as well as the neighborhood chromatic, among other efficient and highly discriminative graph features.

The main contributions are as follows:

- **Edge role discovery:** This work introduces the problem of edge role discovery and proposes a computational framework for learning and modeling edge roles in both static and dynamic networks.
- **Higher-order latent space model:** Introduced a higher-order latent role model that leverages higher-order network features for learning and modeling node and edge roles. We also introduced graphlet-based roles and proposed feature and role learning techniques.
- **Efficient and scalable:** All proposed algorithms are parallelized. Moreover, the feature and role learning and inference algorithms are linear in the number of edges.

2. HIGHER-ORDER EDGE ROLE MODEL

This section introduces our higher-order edge role model and a generalizable framework for computing edge roles based on higher-order network features.

2.1 Initial Higher-order Network Features

Existing role discovery methods use simple degree-based features [16]. In this work, we use graphlet methods [36, 25, 1] for computing higher-order network features based on induced subgraph patterns (instead of simply edge and node patterns) for discovering better and more meaningful roles. Following the idea of feature-based roles [30], we systematically discover an edge-based feature representation. [36, 25] As initial features, we used a recent parallel graphlet decomposition framework proposed in [1] to compute a variety of edge-based graphlet features of size $k = \{3, 4, \ldots\}$. Using these initial features, more representative, explainable,

Table 1: Summary of Bregman Divergences and update rules

	$\phi(y)$	$\nabla^2\phi(y)$	$\mathbb{D}_\phi(x\|x')$	**Update**
Fro.	$y^2/2$	1	$(x-x')^2/2$	$v_{jk} = \frac{\sum_{i=1}^{m} x_{ij}^{(k)} u_{ik}}{\sum_{i=1}^{m} u_{ik} u_{ik}}$
KL	$y \log y$	$1/y$	$x \log \frac{x}{x'} - x + x'$	$v_{jk} = \frac{\sum_{i=1}^{m} x_{ij}^{(k)} u_{ik} / x'_{ij}}{\sum_{i=1}^{m} u_{ik} u_{ik} / x'_{ij}}$
IS	$-\log y$	$1/y^2$	$\frac{x}{x'} - \log \frac{x}{x'}$	$v_{jk} = \frac{\sum_{i=1}^{m} x_{ij}^{(k)} u_{ik} / {x'_{ij}}^2}{\sum_{i=1}^{m} u_{ik} u_{ik} / {x'_{ij}}^2}$

and novel features can be discovered. See the relational feature learning template given in [30]. As an aside, graphlets offer a way to generalize many existing feature learning systems (including those that have been used for learning and extracting roles). We can generalize the above by introducing a generic k-vertex graphlet operator that returns counts and other statistics for any k-vertex induced subgraph where $k > 2$.

2.2 Edge Feature Representation Learning

Learning important and practical representations automatically is useful for many machine learning applications beyond role discovery such as anomaly detection, classification, and descriptive modeling/exploratory analysis. These methods greatly reduce the engineering effort while also revealing important latent features that lead to better predictive performance and power of generalization. This section introduces a generalizable, flexible, and extremely efficient *edge feature learning and inference framework* capable of automatically learning a representative set of edge features automatically. The proposed framework and algorithms that arise from it naturally support arbitrary graphs including undirected, directed, and/or bipartite networks. More importantly, our approach also handles attributed graphs in a natural way, which typically consist of a graph G and a set of arbitrary edge and/or node attributes. The attributes typically represent intrinsic edge and node information such as age, location, gender, political views, textual content of communication between individuals, among other possibilities.

For edge feature learning and extraction, we introduce the notion of an edge neighbor. Intuitively, given an edge $e_i = (v, u) \in E$, let $e_j = (a, b)$ be an edge neighbor of e_i iff $a = v$, $a = u$, $b = v$, or $b = u$. Informally, e_j is a neighbor of e_i if e_j and e_i share a vertex. This definition can easily be extended for incorporating further h-distant neighbors.

The relational operators used to search the space of possible neighbor features at the current and previous learned feature layers include relational operators such as mean, sum, product, min, max, variance, L1, L2, and more generally, any (parameterized) similarity function including positive semidefinite functions such as the Radial Basis Function (RBF) $K\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2)$, polynomial similarity functions of the form $K(\mathbf{x}_i, \mathbf{x}_j) = (a \cdot \langle \mathbf{x}_i, \mathbf{x}_j \rangle + c)^d$, sigmoid neural network kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(a \cdot \langle \mathbf{x}_i, \mathbf{z}_j \rangle + c)$, among others. The flexibility and generalizability of the proposed approach is a key advantage and contribution, *i.e.*, many components are naturally interchangeable, and thus our approach is not restricted to only the relational operators mentioned above, but can easily leverage other application or problem domain specific (relational) operators[1].

Features are searched from the selected feature subspaces and the candidate features that are actually computed are pruned to ensure the set learned is as small and representative as possible capturing novel and useful properties. We define a feature graph G_f where the nodes represent features learned thus far among all current feature layers as well as any initial features which were not immediately pruned due to being redundant or non-informative *w.r.t.* the objective function[2] whereas the edges encode dependencies between the features. Further, the edges are weighted by the computed similarity/correlation (or distance/disagreement) measure, thus as $W_{ij} \to 1$ then the two features f_i and f_j are considered to be extremely similar (and thus possibly redundant), whereas $W_{ij} \to 0$ implies that f_i and f_j are significantly different. In this work, we use log-binning disagreement, though have also used Pearson correlation, among others.

After constructing the weighted feature graph, our approach has two main steps: pruning noisy edges between features and the removal of redundant features (*i.e.*, nodes in G_f). To remove these spurious relationships in the feature graph, we use a simple adaptive sparsification technique. The technique uses a threshold γ which is adapted automatically at each iteration in the search procedure. Once the spurious edges have been removed entirely from the feature graph, we then prune entire features that are found to be redundant. In other words, we discard vertices (i.e., features) from the feature graph that offer no discriminatory power. This can be performed by partitioning the feature graph in some fashion (*e.g.*, connected components). Note that once a feature is added to the set of representative features, it cannot be pruned. However, all features are scored at each iteration, including the representative features, and redundant features are discarded. If a feature is found to closely resemble one of the representative features, we prune it, and keep only the representative feature, as it is more primitive (discovered in a previous iteration). Our approach searches the space of features until one of the following stopping criterion are met: (i) the previous iteration was unfruitful in discovering any novel features, or if (ii) the maximum number of iterations is exceeded which may be defined by the user.

It is worth mentioning that we could have used an arbitrary node feature learning approach such as the one described in [30]. For instance, given a node feature matrix $\mathbf{Z} \in \mathbb{R}^{n \times f}$ (from one such approach), we can easily derive edge features from it (which can then be used for learning edge roles) by using one or more operators over the edges as follows: given an edge $e_k = (v_i, v_j) \in E$ with end points v_i and v_j, one can simply combine the feature values z_i and z_j of v_i and v_j, respectively, in some way, *e.g.*, $x_k = z_i + z_j$ where x_k is the resulting edge feature value for e_k.

2.3 Learning Latent Higher-order Edge Roles

Let $\mathbf{X} = [x_{ij}] \in \mathbb{R}^{m \times f}$ be a matrix with m rows representing edges and f columns representing arbitrary features[3].

[1] See [15] for more details.

[2] In general, the objective function can be either unsupervised or supervised.

[3] For instance, the columns of $\mathbf{X}$ represent arbitrary features such as graph topology features, non-relational features/attributes, and relational neighbor features, among other pos-

More formally, given $\mathbf{X} \in \mathbb{R}^{m \times f}$, the edge role discovery optimization problem is to find $\mathbf{U} \in \mathbb{R}^{m \times r}$ and $\mathbf{V} \in \mathbb{R}^{f \times r}$ where $r \ll \min(m, f)$ such that the product of two lower rank matrices $\mathbf{U}$ and $\mathbf{V}^T$ minimizes the divergence between $\mathbf{X}$ and $\mathbf{X}' = \mathbf{U}\mathbf{V}^T$. Intuitively, $\mathbf{U} \in \mathbb{R}^{m \times r}$ represents the latent *role mixed-memberships* of the edges whereas $\mathbf{V} \in \mathbb{R}^{f \times r}$ represents the contributions of the features with respect to each of the roles. Each row $\mathbf{u}_i^T \in \mathbb{R}^r$ of $\mathbf{U}$ can be interpreted as a low dimensional rank-r embedding of the i^{th} *edge* in $\mathbf{X}$. Alternatively, each row $\mathbf{v}_j^T \in \mathbb{R}^r$ of $\mathbf{V}$ represents a r-dimensional role embedding of the j^{th} feature in $\mathbf{X}$ using the same low rank-r dimensional space. Also, $\mathbf{u}_k \in \mathbb{R}^m$ is the k^{th} column representing a "latent feature" of $\mathbf{U}$ and similarly $\mathbf{v}_k \in \mathbb{R}^f$ is the k^{th} column of $\mathbf{V}$.

For the higher-order latent network model, we solve:

$$\underset{(\mathbf{U},\mathbf{V})\in\mathcal{C}}{\arg\min} \Big\{ \mathbb{D}_\phi(\mathbf{X} \| \, \mathbf{U}\mathbf{V}^T) + \mathcal{R}(U, V) \Big\} \tag{1}$$

where $\mathbb{D}_\phi(\mathbf{X} \| \mathbf{U}\mathbf{V}^T)$ is an arbitrary Bregman divergence [9] between $\mathbf{X}$ and $\mathbf{U}\mathbf{V}^T$. Furthermore, the optimization problem in (1) imposes hard constraints $\mathcal{C}$ on $\mathbf{U}$ and $\mathbf{V}$ such as non-negativity constraints $\mathbf{U}, \mathbf{V} \geq 0$ and $\mathcal{R}(U, V)$ is a regularization penalty. In this work, we mainly focus on solving $\mathbb{D}_\phi(\mathbf{X} \| \mathbf{U}\mathbf{V}^T)$ under non-negativity constraints:

$$\underset{\mathbf{U}\geq 0,\mathbf{V}\geq 0}{\arg\min} \Big\{ \mathbb{D}_\phi(\mathbf{X} \| \, \mathbf{U}\mathbf{V}^T) + \mathcal{R}(U, V) \Big\} \tag{2}$$

Given the edge feature matrix $\mathbf{X} \in \mathbb{R}^{m \times f}$, the edge role discovery problem is to find $\mathbf{U} \in \mathbb{R}^{m \times r}$ and $\mathbf{V} \in \mathbb{R}^{f \times r}$ such that

$$\mathbf{X} \approx \mathbf{X}' = \mathbf{U}\mathbf{V}^T \tag{3}$$

To measure the quality of our edge mixed membership model, we use Bregman divergences:

$$\sum_{ij} \mathbb{D}_\phi(x_{ij} \| x'_{ij}) = \sum_{ij} \big(\phi(x_{ij}) - \phi(x'_{ij}) - \ell(x_{ij}, x'_{ij}) \big)$$

where ϕ is a univariate smooth convex function and

$$\ell(x_{ij}, x'_{ij}) = \nabla\phi(x'_{ij})(x_{ij} - x'_{ij}),$$

where $\nabla^p\phi(x)$ is the p-order derivative operator of ϕ at x. Furthermore, let $\mathbf{X} - \mathbf{U}\mathbf{V}^T = \mathbf{X}^{(k)} - \mathbf{u}_k\mathbf{v}_k^T$ denote the residual term in the approximation (3) where $\mathbf{X}^{(k)}$ is the k-residual matrix defined as:

$$\mathbf{X}^{(k)} = \mathbf{X} - \sum_{h \neq k} \mathbf{u}_h\mathbf{v}_h^T \tag{4}$$

$$= \mathbf{X} - \mathbf{U}\mathbf{V}^T + \mathbf{u}_k\mathbf{v}_k^T, \quad \text{for } k = 1, \ldots, r \tag{5}$$

We use a fast *scalar block coordinate descent approach* that easily generalizes for heterogeneous networks [32]. The approach considers a single element in $\mathbf{U}$ and $\mathbf{V}$ as a block in the block coordinate descent framework. Replacing $\phi(y)$ with the corresponding expression from Table 1 gives rise to a fast algorithm for each Bregman divergence. Table 1 gives the updates for Frobenius norm (Fro.), KL-divergence (KL), and Itakura-Saito divergence (IS). Note that Beta divergence and many others are also easily adapted for our higher-order network modeling framework.

sibilities.

2.4 Model Selection

In this section, we introduce our approach for learning the appropriate model given an arbitrary graph. The approach is leverages the Minimum Description Length (MDL) [14, 29] principle for automatically selecting the "best" higher-order network model. The MDL principle is a practical formalization of Kolmogorov complexity [22]. More formally, the approach finds the model $\mathcal{M}_\star = (\mathbf{V}_r, \mathbf{U}_r)$ that leads to the best compression by solving:

$$M_\star = \underset{M\in\mathcal{M}}{\arg\min} \; \mathcal{L}(M) + \mathcal{L}(\mathbf{X} \,|\, M) \tag{6}$$

where $\mathcal{M}$ is the model space, $M_\star$ is the model given by the solving the above minimization problem, and $\mathcal{L}(M)$ as the number of bits required to encode M using code Ω, which we refer to as the description length of M with respect to Ω. Recall that MDL requires a lossless encoding. Therefore, to reconstruct $\mathbf{X}$ *exactly* from $M = (\mathbf{U}_r, \mathbf{V}_r)$ we must explicitly encode the error $\mathbf{E}$ such that

$$\mathbf{X} = \mathbf{U}_r\mathbf{V}_r^T + \mathbf{E}$$

Hence, the total compressed size of $M = (\mathbf{U}_r, \mathbf{V}_r)$ with $M \in \mathcal{M}$ is simply $\mathcal{L}(X, M) = \mathcal{L}(M) + \mathcal{L}(\mathbf{E})$. Given an arbitrary model $M = (\mathbf{U}_r, \mathbf{V}_r) \in \mathcal{M}$, the description length is decomposed into:

- Bits required to describe the model
- Cost of describing the approximation errors $\mathbf{X} - \mathbf{X}_r = \mathbf{U}_r\mathbf{V}_r^T$ where $\mathbf{X}_r$ is the rank-r approximation of $\mathbf{X}$,

$$\mathbf{U}_r = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_r \end{bmatrix} \in \mathbb{R}^{m \times r}, \quad \text{and} \tag{7}$$

$$\mathbf{V}_r = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_r \end{bmatrix} \in \mathbb{R}^{f \times r} \tag{8}$$

The model $M_\star$ is the model $M \in \mathcal{M}$ that minimizes the total description length: the model description cost X and the cost of correcting the errors of our model. Let $|\mathbf{U}|$ and $|\mathbf{V}|$ denote the number of nonzeros in $\mathbf{U}$ and $\mathbf{V}$, respectively. Thus, the model description cost of M is: $\kappa r(|\mathbf{U}| + |\mathbf{V}|)$ where κ is the bits per value. Similarly, if $\mathbf{U}$ and $\mathbf{V}$ are dense, then the model description cost is simply $\kappa r(m + f)$ where m and f are the number of edges and features, respectively. Assuming errors are non-uniformly distributed, one possibility is to use KL divergence (see Table 1) for the error description cost[4]. The cost of correcting a single element in the approximation is $\mathbb{D}_\phi(x \| x') = x \log \frac{x}{x'} - x + x'$ (assuming KL-divergence), and thus, the total reconstruction cost is:

$$\mathbb{D}_\phi(\mathbf{X} \| \mathbf{X}') = \sum_{ij} X_{ij} \log \frac{X_{ij}}{X'_{ij}} - X_{ij} + X'_{ij} \tag{9}$$

where $\mathbf{X}' = \mathbf{U}\mathbf{V}^T \in \mathbb{R}^{m \times f}$. Other possibilities are given in Table 1. The above assumes a particular representation scheme for encoding the models and data. Recall that the optimal code assigns $\log_2 p_i$ bits to encode a message [34]. Lloyd-Max quantization [26, 23] with Huffman codes [18, 35] are used to compress the model and data [28, 7]. Notice that we require only the length of the description using the above encoding scheme, and thus we do not need to materialize the codes themselves. This leads to the improved model description cost: $\bar{\kappa} r(|\mathbf{U}| + |\mathbf{V}|)$ where $\bar{\kappa}$ is the mean bits required to encode each value[5]. In general, our higher-order

[4]The representation cost of correcting approximation errors

[5]Note $\log_2(m)$ quantization bins are used

network modeling framework can easily leverage other model selection techniques such as AIC [3] and BIC [33].

3. DYNAMIC EDGE ROLE MODEL

This section introduces the *dynamic edge role mixed-membership model* (DERM) and proposes a computational framework for computing edge roles in dynamic networks.

3.1 Dynamic Graph Model & Representation

Given a graph stream $G = (V, E)$ where $E = \{e_1, \ldots, e_m\}$ is an ordered set of edges in the graph stream such that $\tau(e_1) \leq \tau(e_2) \leq \cdots \leq \tau(e_m)$. Note that $\tau(e_i)$ is the edge time for $e_i \in E$ (which may be the edge activation time, arrival time, among other possibilities). Intuitively, E is an infinite edge streaming network where edges arrive continuously over time. From this edge stream, we derive a dynamic network $\mathcal{G} = \{G_t\}_{t=1}^{T}$ where $G_t = (V, E_t)$ represents a snapshot graph at time t. Note that time t is actually a discrete time interval $[a, b)$ where a and b are the start and end time, respectively. Therefore, $E_t = \{e_t \in E \mid a \leq \tau(e_i) < b\}$ and $E = E_1 \cup E_2 \cup \cdots \cup E_T$.

3.2 Dynamic Edge Role Learning

We start by learning a time series of features automatically. Let $G_{1:k} = (V, E_{1:k})$ be the initial dynamic training graph where $E_{1:k} = E_1 \cup \cdots \cup E_k$ and k represents the number of snapshot graphs to use for learning the initial set of (representative) dynamic features. Given $\{G_t\}_{t=1}^{T}$ and $G_{1:k} = (V, E_{1:k})$, the proposed approach automatically learns a set of features $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$ where each $f_i \in \mathcal{F}$ represents a learned feature definition from $G_{1:k}$. Given the learned role definitions $\mathbf{V} \in \mathbb{R}^{r \times d}$ using a subset of past temporal graphs, we then estimate the edge role memberships $\{\mathbf{U}_t\}_{t=1}^{T}$ for each $\{G_t\}_{t=1}^{T}$ (and any future graph snapshots $G_{t+1}, \ldots, G_{t+p}$) where $\mathbf{U}_t \in \mathbb{R}^{m \times r}$ is an edge by role membership matrix. The dynamic edge role model is selected using the approach proposed in Section 2.4.

Time-scale Learning: This section briefly introduces the problem of learning an appropriate time-scale automatically and proposes a few techniques. The time-scale learning problem can be formulated as an optimization problem where the optimal solution is the one that minimizes the objective function. Naturally, the objective function encodes the error from models learned using a particular time-scale s (*e.g.*, 1 minute, 1 hour). Thus solving the optimization problem leads to identifying models from the time-scale s that lead to the least error.

Updating Features and Role Definitions: To prevent the features and role definitions from becoming stale and meaningless over time (due to temporal/concept drift as the network and its attributes/properties evolve), we use the following approach: the loss (or another measure) is computed and tracked over time, and when it becomes too large (from either the features or roles), we then re-compute the feature definitions $\mathcal{F}$ and role definitions. Both the features and roles definitions can be learned in the background as well, and can obviously be computed in parallel. The edge role framework is also flexible for other types of approaches, and thus, not limited to the simple approach above (which is a key advantage of this work).

4. EXPERIMENTS

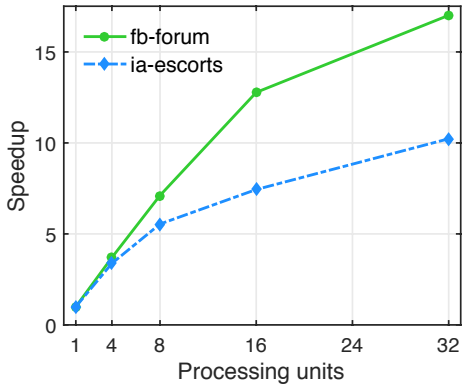


Figure 1: Higher-order role discovery shows strong scaling as we increase the number of processing units.

This section investigates the scalability and effectiveness of the higher-order latent space modeling framework.

Scalability: We investigate the scalability of the parallel framework for modeling higher-order latent edge roles. To evaluate the effectiveness of the parallel modeling framework, we measure the speedup defined as simply $S_p = T_1/T_p$ where T_1 is the execution time of the sequential algorithm, and T_p is the execution time of the parallel algorithm with p processing units. Overall, the methods show strong scaling (See Figure 1). Similar results were observed for other networks. As an aside, the experiments in Figure 1 used a 4-processor Intel Xeon E5-4627 v2 3.3GHz CPU.

Higher-order Model Selection: MDL is used to automatically learn the appropriate edge role model. In Figure 2, description length (in bits) is minimized when $r = 18$. Intuitively, too many roles increases the model description cost, whereas too few roles increases the cost of describing errors. In addition, Figure 3 shows the runtime of our approach. Furthermore, Figure 5 demonstrates the impact on

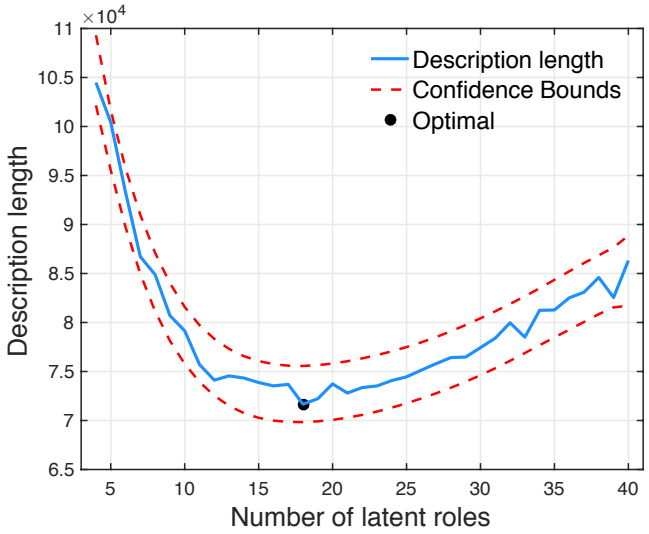


Figure 2: In the example shown, the valley identifies the correct number of latent roles.

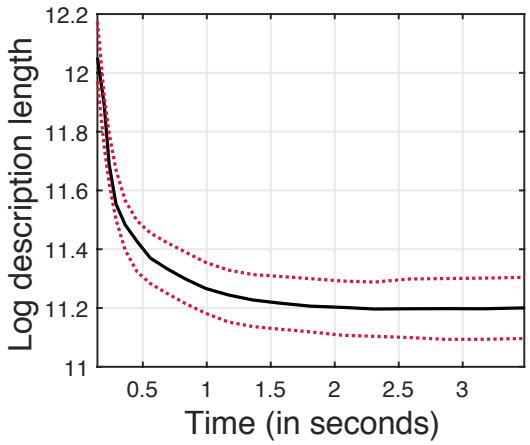


Figure 3: The running time of our approach. The x-axis is time in seconds and the y-axis is the log description cost. The curve is the average over 50 experiments and the dotted lines represent three standard deviations. The result reported above is from a laptop with a single core.

the learning time, number of novel features discovered, and their sparsity, as the tolerance (ε) and bin size (α) varies.

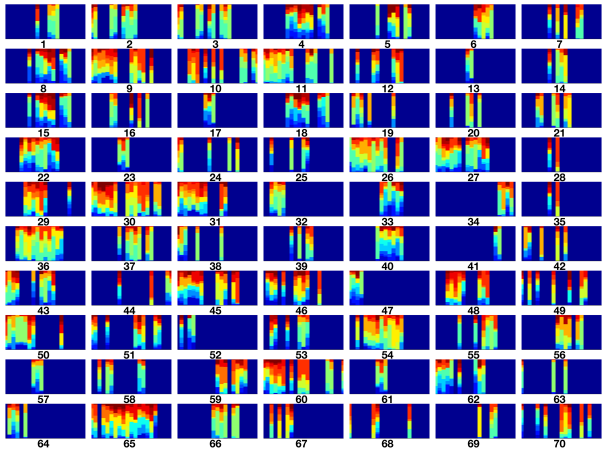
Modeling Dynamic Networks: In this section, we investigate the Enron email communication networks using the *Dynamic Edge Role Mixed-membership Model* (DERM). The Enron email data consists of 151 Enron employees whom have sent 50.5k emails to other Enron employees. We processed all email communications spanning over 3 years of email communications, and discarded the textual content of the email, and only use the edges representing a directed email communication (from one employee to another). The email communications are from 05/11/1999 to 06/21/2002.

For learning edge roles (and a set of representative edge features), we leverage the first year of emails. Note that other work such as dMMSB [12] use email communications from 2001 only, which corresponds to the time period that the Enron scandal was revealed (October 2001). We instead study a much more difficult problem. In particular, given only past data, can we actually uncover and detect the key events leading up to the downfall of Enron? A dynamic network $\{G_t\}_{t=1}^{T}$ is constructed from the remaining email communications (approximately 2 years) where each snapshot graph G_t, $t = 1, \ldots, T$ represents a month of communications. Interestingly, we learn a *dynamic node role mixed-membership model* with 5 latent roles, which is exactly the number of *latent node roles* learned by dMMSB [12]. However, we learn a dynamic *edge* role mixed-membership model with 18 roles. Evolving edge and node mixed-memberships from the Enron email communication network are shown in Figure 4. The set of edges and nodes visualized in Figure 4 are selected using the difference entropy rank (See Eq.(10) below) and correspond to the edges and nodes with largest difference entropy rank $\mathbf{d}$. The first role in Figure 4 represents inactivity (dark blue).

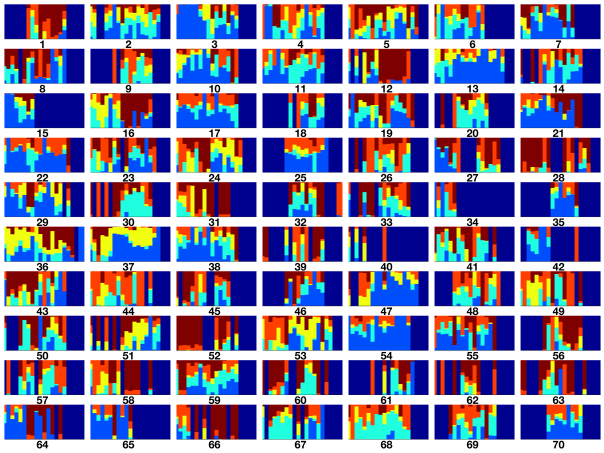
For identifying anomalies, we use the difference entropy rank defined as:

$$\mathbf{d} = \max_{t \in T} H(\mathbf{u}_t) - \min_{t \in T} H(\mathbf{u}_t) \tag{10}$$

where $H(\mathbf{u}_t) = -\mathbf{u}_t \cdot \log(\mathbf{u}_t)$ and $\mathbf{u}_t$ is the r-dimensional



(a) Evolving *edge role* mixed-memberships



(b) Evolving *node role* mixed-membership

Figure 4: Temporal changes in the edge and node mixed-membership vectors (from the Enron email communication network). The horizontal axes of each subplot is time, whereas the vertical axes represent the components of each mixed-membership vector. Roles are represented by different colors.

mixed-membership vector for an edge (or node) at time t. Using the difference entropy rank, we are able to reveal important communications between key players involved in the Enron Scandal, such as Kenneth Lay, Jeffrey Skilling, and Louise Kitchen. Notice that when node roles are used for identifying dynamic anomalies in the graph, we are only provided with potentially malicious employees, whereas using edge roles naturally allow us to not only detect the key malicious individuals involved, but also the important relationships between them, which can be used for further analysis, among other possibilities.

Exploratory Analysis: Figure 6 visualizes the node and edge roles learned for ca-netscience. While our higher-order latent space model learns a stochastic r-dimensional vector for each edge (and/or node) representing the individual role memberships, Figure 6 assigns a single role to each link and node for simplicity. In particular, given an edge $e_i \in E$ (or node) and its mixed-membership row vector $\mathbf{u}_i$, we assign e_i the role with maximum likelihood $k_\star \leftarrow \arg\max_k u_{ik}$. The

t/b	0.5	0.6	0.7	0.8	0.9
0.01	1.48	0.95	0.57	0.47	0.41
0.05	1.03	0.55	0.48	0.46	0.45
0.1	0.72	0.57	0.54	0.51	0.48
0.2	0.78	0.58	0.55	0.52	0.49
0.5	0.58	0.56	0.54	0.6	0.56

(a) Learning time

t/b	0.5	0.6	0.7	0.8	0.9
0.01	327	149	81	46	26
0.05	168	73	48	31	18
0.1	111	53	42	26	18
0.2	94	49	36	24	18
0.5	39	33	30	21	16

(b) Number of features discovered

t/b	0.5	0.6	0.7	0.8	0.9
0.01	0.151	0.158	0.136	0.097	0.077
0.05	0.23	0.209	0.169	0.111	0.084
0.1	0.235	0.23	0.186	0.133	0.084
0.2	0.24	0.223	0.222	0.143	0.084
0.5	0.319	0.276	0.242	0.158	0.094

(c) Sparsity of features

Figure 5: Impact on the learning time, number of features, and their sparsity, as the tolerance (ε) and bin size (α) varies.

higher-order edge and node roles from Figure 6 are clearly meaningful. For instance, the red edge role represents a type of bridge relationship as shown in Figure 6.

Sparse Graph Feature Learning: Recall that the proposed feature learning approach attempts to learn "sparse graph features" to improve learning and efficiency, especially in terms of space-efficiency. This section investigates the effectiveness of our sparse graph feature learning approach. Results are presented in Table 2. In all cases, our approach learns a highly compressed representation of the graph, requiring only a fraction of the space of current (node) approaches. Moreover, the density of edge and node feature representations learned by our approach is between $[0.164, 0.318]$ and $[0.162, 0.334]$ for nodes (See $\rho(\mathbf{X})$ and $\rho(\mathbf{Z})$ in Table 2) and up to $6x$ more space-efficient than other approaches. While existing feature learning approaches for graphs are unable to learn higher-order graph features (and thus impractical for higher-order network analysis and modeling), they also have another fundamental disadvantage: they return dense features. Learning space-efficient features is critical especially for large networks. For instance, notice that on extremely large networks, storing even a small number of edge (or node) features quickly becomes impractical. Despite the importance of learning sparse graph features, existing work has ignored this problem as most approaches stem from Statistical Relational Learning (SRL) [13] and have been designed for extremely small graphs. Moreover, nearly all existing methods focus on node features [10, 19, 21, 20], whereas we focus on both and primarily on learning novel and important edge feature representations from large massive networks.

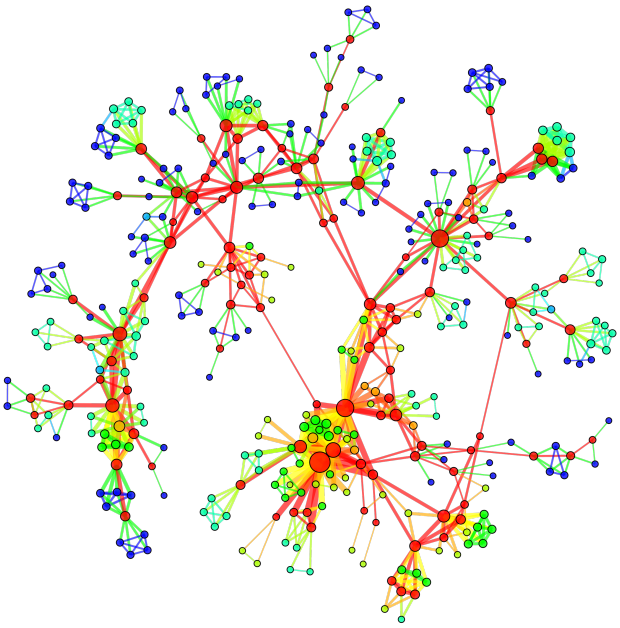


Figure 6: Edge and node roles for ca-netscience. Link color represents the edge role and node color indicates the corresponding node role.

Table 2: Higher-order sparse graph feature learning for latent node and edge network modeling. Recall that f is the number of features, L is the number of layers, and $\rho(\mathbf{X})$ is the sparsity of the feature matrix. Edge values are bold.

graph	f		L		$\rho(\mathbf{X})$	$\rho(\mathbf{Z})$
socfb-MIT	**2080**	(912)	**8**	(9)	**0.318**	(0.334)
yahoo-msg	**1488**	(405)	**7**	(7)	**0.164**	(0.181)
enron	**843**	(109)	**5**	(4)	**0.312**	(0.320)
Facebook	**1033**	(136)	**7**	(5)	**0.187**	(0.162)
bio-DD21	**379**	(723)	**6**	(6)	**0.215**	(0.260)

Computational Complexity: Recall that m is the number of edges, n is the number of nodes, f is the number of features, and r is the number of latent roles. The total time complexity of the *higher-order latent space model* is: $\mathcal{O}\big(f(m+nr)\big)$. Thus, the runtime is linear in the number of edges. The time complexity is decomposed into the following main parts: Feature learning takes $\mathcal{O}(f(m+nf))$. Model learning takes $\mathcal{O}(mrf)$ in the worst case (which arises when $\mathbf{U}$ and $\mathbf{V}$ are completely dense). The quantization and Huffman coding terms are very small and therefore ignored. Latent role learning using scalar element-wise coordinate descent has worst case complexity of $\mathcal{O}(mfr)$ per iteration which arises when $\mathbf{X}$ is completely dense. However, assuming $\mathbf{X}$ is sparse, then it takes $\mathcal{O}(|\mathbf{X}|\, r)$ per iteration where $|\mathbf{X}|$ is the number nonzeros in $\mathbf{X} \in \mathbb{R}^{m \times f}$. In addition, we compute the initial set of graphlet-based features using the efficient parallel algorithm in [2]. Note that this algorithm computes the counts of a few graphlets and directly obtain the others in constant time. This takes $\mathcal{O}(\Delta\big(|S_u|+|S_v|+|T_e|\big))$ for any given edge $e_i = (v, u)$, where Δ is the maximum degree for any vertex, S_v, S_u are the sets of wedge nodes and T_e is the set of triangles incident to edge e_i.

5. CONCLUSION

This work introduced the notion of edge roles and proposed

a higher-order latent space network model for edge role discovery. To the best of our knowledge, this work is the first to explore using higher-order graphlet-based features for role discovery. Moreover, these features are counts of various induced subgraphs of arbitrary size and were used directly for role discovery as well as given as input into a graph representation learning approach to learn more discriminative features based on these initial features. Furthermore, feature-based edge roles also have many important and key properties and can be used for graph similarity, node and edge similarity queries, visualization, anomaly detection, classification, link prediction, among many other tasks. Our edge role discovery framework also naturally supports large-scale attributed networks.

6. REFERENCES

[1] N. K. Ahmed, J. Neville, R. A. Rossi, and N. Duffield. Efficient graphlet counting for large networks. In *ICDM*, page 10, 2015.

[2] N. K. Ahmed, J. Neville, R. A. Rossi, N. Duffield, and T. L. Willke. Graphlet decomposition: Framework, algorithms, and applications. *Knowledge and Information Systems (KAIS)*, pages 1–32, 2016.

[3] H. Akaike. A new look at the statistical model identification. *Transactions on Automatic Control*, 19(6):716–723, 1974.

[4] C. Anderson, S. Wasserman, and K. Faust. Building stochastic blockmodels. *Social Networks*, 14(1):137–161, 1992.

[5] P. Arabie, S. Boorman, and P. Levitt. Constructing blockmodels: How and why. *Journal of Mathematical Psychology*, 17(1):21–63, 1978.

[6] V. Batagelj, A. Mrvar, A. Ferligoj, and P. Doreian. Generalized blockmodeling with pajek. *Metodoloski zvezki*, 1:455–467, 2004.

[7] W. R. Bennett. Spectra of quantized signals. *Bell System Technical Journal*, 27(3):446–472, 1948.

[8] S. Borgatti, M. Everett, and J. Johnson. *Analyzing Social Networks*. Sage Publications, 2013.

[9] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Math. and Mathematical Physics*, 7(3):200–217, 1967.

[10] J. Davis, I. M. Ong, J. Struyf, E. S. Burnside, D. Page, and V. S. Costa. Change of representation for statistical relational learning. In *IJCAI*, pages 2719–2726, 2007.

[11] P. Doreian, V. Batagelj, and A. Ferligoj. *Generalized Blockmodeling*, volume 25. Cambridge University Press, 2005.

[12] W. Fu, L. Song, and E. P. Xing. Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 329–336, 2009.

[13] L. Getoor and B. Taskar, editors. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[14] P. D. Grünwald. *The minimum description length principle*. MIT press, 2007.

[15] I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh. *Feature Extraction: Founds and Applications*. Springer, 2008.

[16] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li. Rolx: Structural role extraction & mining in large graphs. In *SIGKDD*, pages 1231–1239, 2012.

[17] P. HollandKathryn Blackmond and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[18] D. A. Huffman et al. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[19] S. Kok and P. Domingos. Statistical predicate invention. In *ICML*, pages 433–440. ACM, 2007.

[20] N. Landwehr, K. Kersting, and L. De Raedt. nFOIL: Integrating naıve bayes and FOIL. In *AAAI*, pages 795–800, 2005.

[21] N. Landwehr, A. Passerini, L. De Raedt, and P. Frasconi. kfoil: Learning simple relational kernels. In *AAAI*, volume 6, pages 389–394, 2006.

[22] M. Li and P. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer Science & Business Media, 2009.

[23] S. Lloyd. Least squares quantization in pcm. *Transactions on Information Theory*, 28(2):129–137, 1982.

[24] F. Lorrain and H. White. Structural equivalence of individuals in social networksâĂă. *Journal of Mathematical Sociology*, 1(1):49–80, 1971.

[25] D. Marcus and Y. Shavitt. Rage–a rapid graphlet enumerator for large networks. *Computer Networks*, 56(2):810–819, 2012.

[26] J. Max. Quantizing for minimum distortion. *Transactions on Information Theory*, 6(1):7–12, 1960.

[27] K. Nowicki and T. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[28] B. Oliver, J. Pierce, and C. E. Shannon. The philosophy of pcm. *Proceedings of the IRE*, 36(11):1324–1331, 1948.

[29] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[30] R. A. Rossi and N. K. Ahmed. Role discovery in networks. *TKDE*, 27(4):1112–1131, 2015.

[31] R. A. Rossi, B. Gallagher, J. Neville, and K. Henderson. Modeling dynamic behavior in large evolving graphs. In *WSDM*, pages 667–676, 2013.

[32] R. A. Rossi and R. Zhou. Parallel Collective Factorization for Modeling Large Heterogeneous Networks. In *Social Network Analysis and Mining*, page 30, 2016.

[33] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[34] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(1):623âĂŞ656, 1948.

[35] J. Van Leeuwen. On the construction of huffman trees. In *ICALP*, pages 382–410, 1976.

[36] S. Wernicke and F. Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.

[37] D. White and K. Reitz. Graph and semigroup homomorphisms on networks of relations. *Social Networks*, 5(2):193–234, 1983.