Investigating Microstructure Patterns of Enterprise Network in Perspective of Ego Network

Xiutao Shi[1], Liqiang Wang[1], Shijun Liu[1,2], Yafang Wang[1], Li Pan[1,2], and Lei Wu[1,2]

[1]School of Computer Science and Technology, Shandong University, Jinan 250101, China
[2]Engineering Research Center of Digital Media Technology, Ministry of Education, Jinan 250101, China
{shixtwy,wanglq1989}@163.com
{lsj,yafang.wang,panli,i_lily}@sdu.edu.cn

Abstract. In social networks the behavior of individuals can be researched through the evolution of the microstructure. As we know, triad is the basic atom shape to build the whole social network. However we find that quad plays the basic role rather than triad in Enterprise Network (EN). In particular, we focus on four typical microstructure patterns including triad, 4-cycle, 4-chordalcycle and 4-clique in EN. We propose algorithms to mine these microstructure patterns and compute the frequencies of each type of microstructure patterns in an efficient parallel way. We also analyze the structural features of these microstructure patterns in a perspective of ego network. Additionally we present the evolutionary rules between these microstructure patterns based on the statistical analysis. Finally we combine the features into traditional methods to solve the link prediction problem. The results show that these features and our combination methods are effective to predict links between enterprises in EN.

Keywords: Microstructure Patterns; Enterprise Network; Ego Network; Triad; Quad; Evolutionary Rules; Link Prediction

1 Introduction

Motivation. Nowadays, the links between enterprises are becoming tighter and the inter-enterprise relations are getting more and more complex. These links and relations are producing huge amounts of data, e.g., the information of the enterprise, the business interactions between enterprises and the unstructured data on social business. One of the most valuable data is the large enterprise network (EN).

In the social sciences, social structure is the patterned social arrangements in society [1]. These structures are emergent from the actions of the individuals, thus we can analyze the features of individual actions through the social structures. The microstructures in EN can not only reflect the behavior of a single enterprise, but also reflect the way and characteristics of interactions between enterprises. It is possible to study the interaction patterns among enterprises and make more accurate recommendations and predictions based on these features. Researchers have studied enterprise networks at multiple levels of structures, including the dyad [2], the ego network [3], and the overall network [15]. However, less attention has been paid to triads and quads yet.

Contributions. We propose a parallel algorithm to search both triad and quad graphlets [5] (inc. 4-cycle, 4-chordalcycle and 4-clique) in EN, which reduced time complexity compared to other approaches. We conduct experiments to study and evaluate triad and quad graphlets to analyze the behaviors of different enterprises. We also apply triad and quad features for link prediction, and the results demonstrate that the methods with quad features performs better than the other features in EN.

2 Related Work

Microstructures Analysis. While analysing microstructures of 3-nodes and 4-nodes, Ahmed et al.[4] proposed a fast, efficient and parallel algorithm for counting microstructures of size k=3, 4-nodes. Trpevski et al.[17] found that vertex signature vector and microstructure correlation matrix are powerful tools for network analysis. Yanardag and Vishwanathan [20] presented a framework to learn latent representations of substructures for graphs. It leveraged the dependency information between sub-structures to improve classification accuracy. Madhavan, Gnyawali and He [15] found a characteristic of enterprise networks: enterprises tend to form triads defined by geography. This is a situation similar to network closure. Biswas et al.[7] focused on ego centric community detection in network data, mainly emphasizing on structural aspects, i.e., reachability and isolability. Dunbar et al.[8] studied the internal structure of these networks to determine whether they have the same kind of layered structure as offline face-to-face networks. Toral et al.[16] analyzed the social network structures in both macro and micro view. Li and Daie [13] proposed a hierarchical clustering method for configuring assembly supply chains to evaluate the coupling according to the product variety information. Girard et al.[10] studied how students social networks emerge by documenting systematic patterns in the process of friendship formation of incoming students. The shape of local and global network structures resulting from this process. Gordon and McCann [11] distinguished three ideal-typical models of processes which may underlie spatial concentrations of related activities. Survey data is used for the London conurbation to explore the relations between concentration and different forms of linkage. These works are mostly focus on analyzing features and evolution of triad. In this paper, we investigated quads in EN to analyze the behaviors between enterprises.

Link Prediction. Lou et al.[14] employed graphical model to predict reciprocity and triadic closure. Huang et al.[12] studied the triadic closure information in dynamic social network and proposed a probabilistic factor model for modeling and prediction. Dong et al.[9] predicted links in heterogeneous networks by a ranking factor graph model. Bhuiyan et al.[6] proposed GUISE, which uses a Markov Chain Monte Carlo (MCMC) sampling method for constructing the approximate Graphlet Frequency Distribution (GFD) of a large network. Wasserman and Pattison [19] described a series of models for investigating the structures in social networks, including several generalized stochastic block models. This paper evaluated the effectiveness of microstructure patterns for link prediction in EN.

3 Mining Microstructure Pattern in EN

3.1 Definitions

Enterprise Network (EN). As defined, EN = (ENT, E). ENT is the enterprise node collection, $i \in$ ENT, i is an enterprise and E is the edge collection. $e \in E$, $e = (i, j)$, i is a supplier and j is a manufacturer, the edge is from i to j [18].

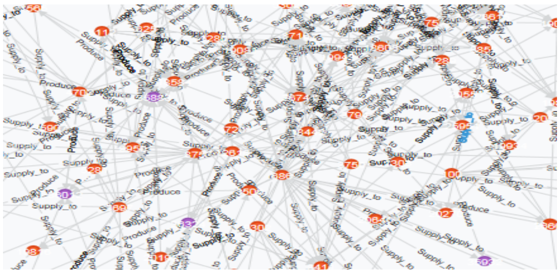


Fig. 1: *The visualization of a part of EN.*

Ego Network. Researchers usually examine the business patterns in the ego network with a central enterprise, for example, manufacturers concern more about their direct-connected suppliers, so we focus on the microstructures in the perspective of ego network.

Ego network is a well-known social phenomenon which deals with individual interest and the relationship [7]. Ego networks consist of a central node (ego) and the nodes to whom ego is directly connected (these are called alters). We can choose an arbitrary node to be the ego and different egos lead to different types of ego networks. For example, take a manufacturer as the ego, the manufacturer and its suppliers form an ego network. This ego network has different characteristics with the ego network that has a supplier ego.

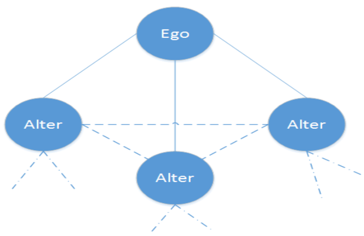
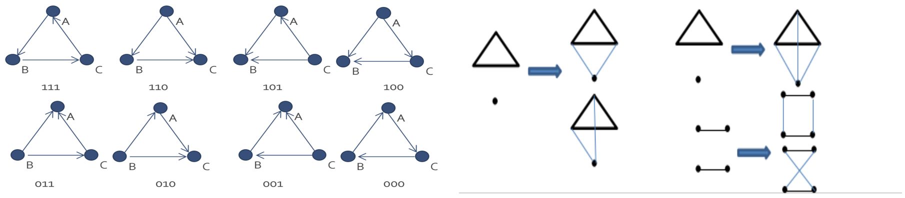


Fig. 2: *Example of ego network. Any arbitrary central node (Ego) which connects other nodes (Alters) in the network (solid lines) and connectivity among alters (dashed lines). Connectivity with rest of the network is represented with dashed-dotted lines.*

3.2 Category of Microstructures

Among the multiple levels of microstructure, triad−subsets of three network nodes and the possible ties among them–are the important topic of early research on social networks. However triads are not so valuable in EN. In a real business scenario, two enterprises A and B usually interact with each other directly rather than through other enterprises. Quad contains more information than triad including both indirect and direct relationships between two connected enterprises. Therefore, quads are more important than triads in EN. As EN is directional, there are more kinds of triads and quads compared with the undirected graph. Additionally, there is no two-way relationship between enterprises in this directed network, so there are only 8 possible different closed triads.



(a) $Triads$ (b) $Quads$

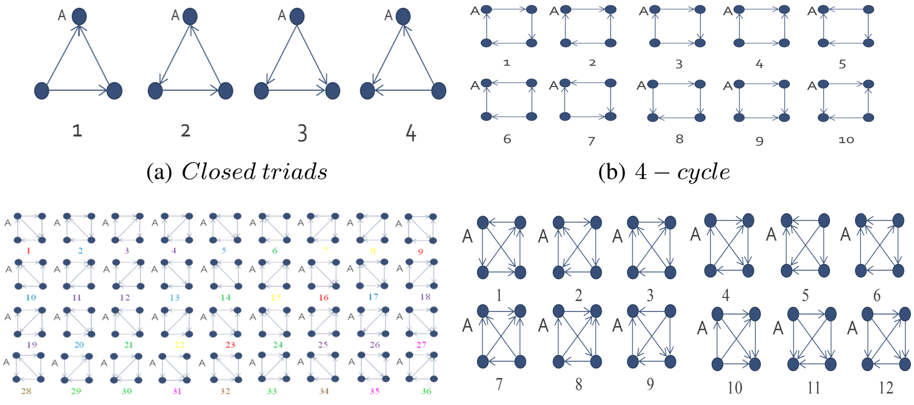
Fig. 3: *Triad code representation and the formation of quads.*

As shown in Figure 3(a), we encode every triad with a 3-length binary number. A, B and C represent the enterprise. We defined that $A \rightarrow B$, $B \rightarrow C$ and $C \rightarrow A$ is positive direction represented by 1. The relationship AB represents that enterprise A supplies products to enterprise B. Meanwhile, 0 means negative direction. For example, 111 represents one type of triad that contains $A \rightarrow B$, $B \rightarrow C$ and $C \rightarrow A$. Figure 3(b) shows the forming of several critical microstructures, which works as the basic rule of our closed quads search algorithm. A closed triad and a point associated with it will form a 4-clique or 4-chordalcycle [4]. Two completely different relationships form a 4-cycle [4] by linking the corresponding points together. However, as the EN is directed, a structural classification for closed triads and the three classical quads are the main features used in our methods.

3.3 The Patterns of Triads and Quads

The closed triads and three representative kinds of closed quads are classified by relations between alters and ego. We need to figure out the specific classification and the mentioned structures in each category. In this way, we can not only calculate the frequency of different types of graphlets that each enterprise has, but also generalize the isomorphic patterns. We show the classifications in Figure 4.

Obviously, triad has 4 possible 3-nodes variants. Subsequently, we get the data statistics (shown in Figure 5) for involved enterprises in terms of this classification. Some triads (shown in Figure 4(a)) can be categorized into these four variants by rotating around A.



(a) $Closed\ triads$ (b) $4-cycle$

(c) $4-chordalcycle$ (d) $4-clique$

Fig. 4: *Closed triads and quads are classified by ego network. Node labeled by A represents the ego, other nodes represent alters. These triads and quads are classified according to the ego network of A.*

For example, the triad 110 becomes triad 100 by rotation and both of them belong to triad No.3 shown in Figure 4(a). Similarly, some 4-nodes graphlets can also be converted into the same type shown. For example, 4-cycle 0000 and 4-cycle 1111 belong to 4-cycle No.1 shown in Figure 4(b). The 4-clique $111111, 110001, 101111, 110000, 001010$ and 000010 all belong to 4-clique No.1 shown in Figure 4(d), and 4-chordalcycle $000012, 111112$ belong to 4-chordalcycle No.1 shown in Figure 4(c).

3.4 Mining Algorithms

In this section, we describe our approach for querying and recording the microstructure patterns. This algorithm takes only a fraction of the time compared with the prevailing methods. Given the enterprise graph $G = (V, E)$. Firstly, our proposed algorithm traverses all the edges when querying the triad. We define that $N(u)$ and $N(v)$ are the set of neighbors of u and v respectively. Given a single edge $e = (u, v) \in E$, if a node $w \in N(u) \cap N(v)$, we record u, v and w with their information and relationship between them, then we get a triad. Based on these triads, we start to search 4-nodes microstructures (Quads). As shown in Figure 3(b), the searching principle is that the 4-chordalcycle and 4-clique can be decomposed into 3-nodes triads and the 4-cycle can be decomposed into two edges. We summarize this procedure in the following steps:
STEP 1: For each edge $e = (u, v)$, find all the common neighborhood nodes of u and v to form closed triad, and record their information and relationships between them.
STEP 2: For each closed triad, find all possible nodes which can form the 4-chordalcycle or 4-clique, and record the information of nodes and relationships.
STEP 3: For each edge $e1 = (u1, v1)$, find another edge $e2 = (u2, v2)$ to form the 4-cycle, and record the information of this 4-cycle.

STEP 4: For all closed triad and 4-nodes quads, divide them into different types and analyze them.

Algorithm 1 : Mining Triad Graphlets

Input: $EN\ G = (V, E)$
Output: $TRIAD$.
1: **for** $edge\ e = (u, v) \in E$ **do**
2: **if** u==v **then**
3: then continue;
4: $Union = N(u) \cap N(v)$;
5: **for** $w \in Union$ **do**
6: $recordToTriad(u, v, w, GetNodeRelation(u, v, w))$;

Lines 1-5 of algorithm 1 show how to find and record triads in EN. Only when node w has relationships with both u and v in edge $e = (u, v)$, the pattern (u, v, w) could be a closed triad. However, the additional relevant functions, such as the approach to record the microstructures or remove the duplicates, are beyond our discussions, as they should be designed specifically for a concrete requirement.

Lines 1-9 of algorithm 2 show how to find and record 4-clique and 4-chordalcycle. For any node w, w could form a 4-clique or 4-chordalcycle with triad t, if and only if $|N(w)|$ is not less than 2. We define $Linkpoints$ as the set of overlapping nodes in the neighborhoods of w and t, and define $TRIAD$ as the set of all triads in EN and $QUAD$ as the set of all quads. If t is completely in the ego network of w, it illustrates that every node has links with each other and these four nodes could form a 4-clique. And if the number of $Linkpoints$ is 2, it means that two nodes in t have relationships with w and these four nodes could be a 4-chordalcycle. Then we just need to select and record the information of these nodes and relationships, such as ID, direction of relationships. Lines 10-17 describe the process of finding and recording the 4-cycle. For every edge $e1 = (u1, v1)$, we try to find a related $e2 = (u2, v2)$ to form a 4-cycle. A feasible $e2$ requires that $e3 = (u1, u2)$ and $e4 = (v1, v2)$ exist while $u1, v1$ are not the neighbor of $v2, u2$ respectively, or similarly $e3 = (u1, v2)$ and $e4 = (u2, v1)$exist while $u1, v1$ are not the neighbor of $u2, v2$ respectively. Of course, if any two nodes are the same in $e1$ and $e2$, they cant form a 4-cycle.

4 Statistic Analysis of Ego Microstructures on the Entire Network

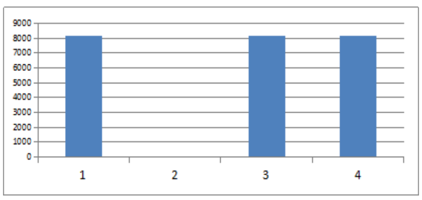
4.1 Ego Triads and Quads Characteristics in EN

Firstly, in Figure 5(a), we can see that the frequency of type 2 is 0 and the frequency of other types is the same. Because there is no transitive supply in enterprise network, that is to say, there is no three enterprises A,B,C existing where A supplies B, B supplies C and C supplies A. The frequencies of other types are the same. The reason is that two different related suppliers supply to the same manufacturer. Secondly, Figure 5(b) shows that type 7 and type 10 are the most commonly occurring 4-cycle. Both 4-cycle No.7

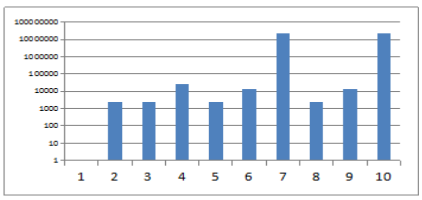
Algorithm 2 : Mining Quad Graphlets.

Input: EN,$G = (V, E)$,$TRIAD$
Output: $QUAD$.

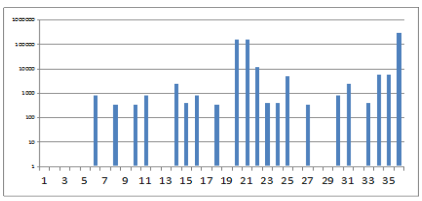
//Find 4-Clique And 4-Chordalcycle
for $node\ w \in V$ **do**
  if $|N(w)| < 2$ **then**
    then continue;
  for $triad\ t \in TRIAD$ **do**
    if $w \in t$ **then**
      then continue;
    if $t \in N(w)$ **then**
      $recordTo4clique(w, t, GetRelation(w, t))$
    else
      $Linkpoints = N(w) \cap t$
    if $|Linkpoints| == 2$ **then**
      $recordTo4chordalcycle(w, Linkpoints, GetRelation(w, Linkpoints)$
// Find 4-Cycle
for $edge\ e1 = (u1, v1) \in E$ **do**
  if $u1 == v1$ **then**
    then continue;
  for $edge\ e2 = (u2, v2) \in E$ **do**
    if $e1 == e2\ or\ AnyPointSame(e1, e2)$ **then**
      then continue;
    if $u1 \in N(u2)\ and\ v1 \in N(v2) and\ u1 \notin N(v2)\ and\ v1 \notin N(u2)$ **then**
      $recordTo4cycle(e1, e2, GetEdgeRelation(e1, e2))$
    if $u1 \in N(v1)\ and\ v1 \in N(u2)\ and\ u1 \notin N(u2)\ and\ v1 \notin N(v2)$ **then**
      $recordTo4chordalcycle(w, Linkpoints, GetRelation(w, Linkpoints)$



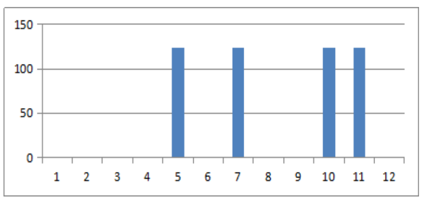
(a) $Closed\ triad$



(b) $4 - cycle$



(c) $4 - chordalcycle$



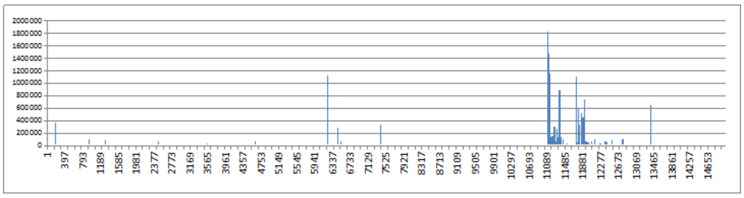
(d) $4 - clique$

Fig. 5: *Number on the X-axis means different types of triad or quad that corresponding to Figure 4. The frequencies of these types are represented by number on the Y-axis.*

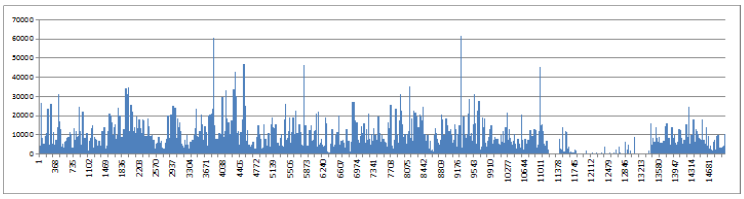
and No.10 are characterized by the distinct division resulting in two manufacturers and two suppliers, which share the manufacturers. As we can see, 4-cycle, 4-chordalcycle and 4-clique have structural dependence with each other. Therefore, by investigating the structures of highest frequency among the 4-chordalcycle, shown in Figure 5(c), we discover the most important feature of these types of 4-chordalcycle compared with 4-cycle: there is a relationship between two manufacturers. Finally, Figure 5(d) shows the number of different types of 4-clique. There is no circular supply chain, and 4-cliques whose frequency are 124 are isomorphic, which means that they share the same structure. In this structure it has a top manufacturer, two related tier suppliers, and a bottom supplier, where the bottom supplier not only indirectly supply to the top manufacturer through two tier suppliers but also directly supply to the top manufacturer.

4.2 Ego Triads and Quads Characteristics between Different Enterprise Types

As mentioned above, 4-cycle No.7 and No.10 are isomorphic and their frequencies are equal, while the distribution of the frequencies of 4-cycle No.7 and No.10 that every enterprise contains varies. Figure 6(b) shows that these enterprises have roughly equal frequency of 4-cycle No.10. As for the distribution of 4-cycle No.7 in Figure 6(a) , the frequencies in several enterprises are typically high while the rest are relatively small, even as low as 0. It also reflects the fact that the number of suppliers is much larger than the number of manufacturers, and the number of suppliers that supplies to large manufacturers is huge compared to the number of manufacturers that large suppliers supply.



(a) type 7 of 4-cycle



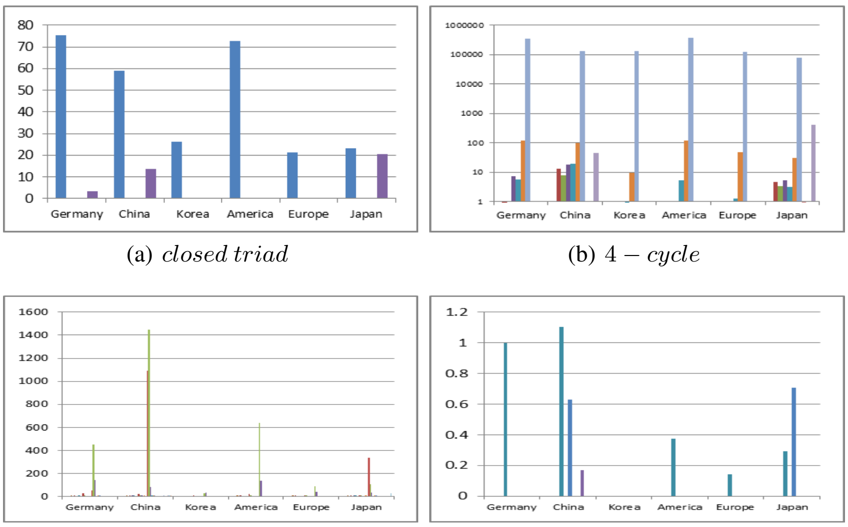
(b) type 10 of 4-cycle

Fig. 6: *Number on the X-axis means ID of these enterprises in this network. Number on the Y-axis means the number of type 7 or 10 of 4-cycle, and these enterprises serve as point A shown in Figure 5(b).*

4.3 Analysis of Automobile Manufacturers

Analysis of Automobile Manufacturers in Different Regions. As the architectures of automobile manufacturing industry in different regions are probably different, we select a number of famous enterprises from the whole network to study the differences in microstructures. Through the result, we can discover the latent features underlying the enterprise network. Figure 7 shows the average quantitative distribution of these enterprises in different microstructures.

From the Figure 7(a), we can see that the enterprises from Korea, America and Europe only appear as manufacturers. Most German enterprises are manufacturers and few German enterprises are intermediaries. Enterprises in China and Japan are not only manufacturers but also intermediaries. From the other figures, most enterprises in different regions play the role of first-tier manufacturer that suppliers directly in microstructures of 4-nodes, and some of these enterprises also serve as the second-tier manufacturers that suppliers indirectly by supplying other enterprises. At the same time, some Chinese and Japanese enterprises are intermediaries, through which suppliers supply manufacturers indirectly.



(a) $closed\ triad$

(b) $4 - cycle$

(c) $4 - chordalcycle$

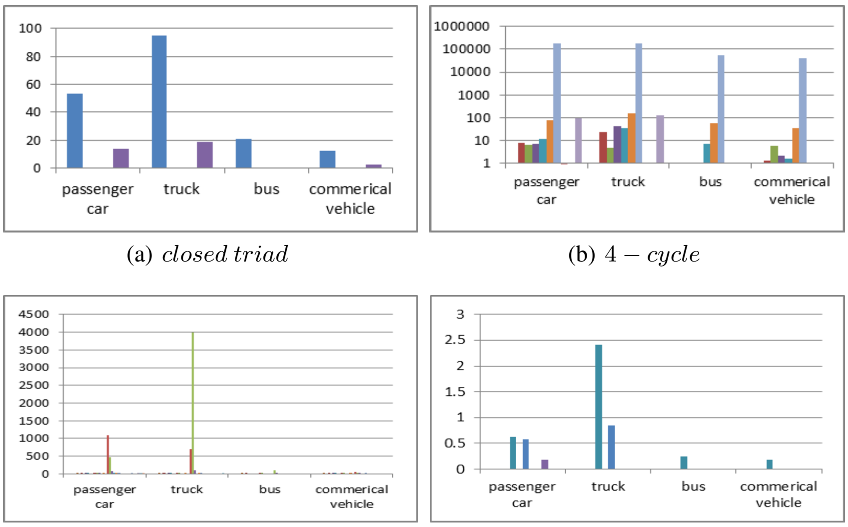
(d) $4 - clique$

Fig. 7: *X-axis: different countries and regions. The number on the Y-axis means the average frequencies of different types of triad or quad. It should be noted that when we count the frequency of triads or quads about one enterprise, this enterprise must be the point A shown in Figure 5.*

Analysis of Automobile Manufacturers Producing Different Cars. The supply network of automobile enterprises that produce different kinds of cars may be different.

Figure 8 shows the average distributions of enterprise that produce passenger cars, trucks, buses and commercial vehicles.

Form Figure 8(a), we can find that these enterprises are mainly manufacturers except bus manufacturers which also appear as intermediaries in network. In addition, few passenger car manufacturers serve as suppliers in enterprise supply network. From the other figures, most of all these enterprises play the role of first tier manufacturer that suppliers directly supply in microstructures of 4-nodes and some of these enterprises also serve as the second tier manufacturers. Moreover, the passenger car manufacturers and truck manufacturers sometimes supply other enterprises directly. However, there is no enterprise that supplies an enterprise indirectly by another enterprise.



(a) $closed\ triad$ (b) $4-cycle$

(c) $4-chordalcycle$ (d) $4-clique$

Fig. 8: *X-axis: different motorcycle type for different purposes. The number on the Y-axis means the average frequencies of different types of triad or quad. It should be noted that when we count the frequencies of triads or quads about one enterprise, this enterprise must be the point A shown in Figure 5.*

5 Microstructure Evolution and Link Predictions

5.1 Evolution between Microstructure Patterns

In this section, we would focus on network evolution among four involved structures, analyzing four processes: 4-cycle evolves into 4-chordalcycle, 4-chordacycle evolves into 4-clique, triad evolves into 4-chordalcycle and triad evolves into 4-clique.

In fact, type 1 to 4 of 4-chordalcycle could be constructed by attaching a bevel edge on type 1 of 4-cycle, which means that there exists a circular supply chain and a relationship where a supplier A supply to C indirectly through B or D. Because the frequency of 4-cycle 1 is 0, the frequencies of 4-chordalcycle No.1-NO.4 are also 0 shown in Figure 5. However, frequencies of some types in 4-chordalcycle are 0 except 4-chordalcycle No.1-No.4. The reason is that these types of 4-chordalcycle contain triad No.2. The most frequently appearing 4-chordalcycle structures: type 36, 20 and 21, could be regarded as 4-cycle No.7 or No.10 added with a supply relationship between the manufacturers. The types 22, 34 and 36 similarly evolves from 4-cycle No.7 or 4-cycle No.10 but added with a supply relationship between the suppliers, rarely appear in Figure 5(c). It demonstrates that two manufacturers are more likely to build a supply relationship between them than suppliers in enterprise network. At the same time, it explains why the frequency of triad is less than 9000 while the frequencies of 4-cycle No.7 and No.10 are more than twenty million: it is rare for two suppliers supplying to the same enterprise to build a supply relationship. However, there is a good chance for two enterprises with the same supplier to establish a relationship in enterprise network. We draw the inferences by observing and analyzing the type 36, 34 and 35 of 4-chordalcycle.

As we can see, there are only types 5, 7, 10 and 11 of 4-clique whose frequencies are not 0. In fact, these four types based on 4-chordalcycle No.20 or No.21 are isomorphic and they all mean that two manufacturers share two suppliers. In addition, a supply relationship exists in two suppliers (or two manufacturers). When a supply relationship is generated between enterprise B and D, this 4-chordalcycle become a 4-clique.

It is not difficult to find that every 4-chordalcycle contains two triads, thus 4-chordalcycle has all the features of triad. If any type of 4-chordalcycle contains the type 2 of triad, the frequency of it must be 0, as shown in Figure 5(c). Similar to 4-chordalcycle, every 4-clique can be viewed as a structure combined with four triads and thus the frequencies of types contained triad No.2 in 4-clique is 0, shown in Figure 5. There are not too many 4-cliques in enterprise network because there is no relationship between two suppliers that supply directly to the same enterprise.

5.2 Link Predictions Model based on Microstructure Patterns

We use closed triad, 4-cycle, 4-chordalcycle and 4-clique as features to improve the prediction accuracy of different algorithms (PMF [22], GPLVM [23], SIMRANK [24]). The goals of these algorithms are predicting whether a relationship will appear between two enterprises. *EE, ET, EF* and *ETF* are used to help prediction because they contain the relationship between enterprises and different kind of microstructures. We carried out a set of experiments to evaluate the performance with *EE, ET, EF* and *ETF*. The definition of *EE, ET, EF* and *ETF* are as follow.

(1) *EE* represents algorithms predicting the relationship by using only the enterprise-enterprise relationship network (Matrix $\mathbf{EE}$) without those microstructures,

(2) *ET* means that algorithm uses the relationship between enterprise and closed triads No.1-No.4 (shown in Figure 4) for predictions.

(3) *EF* means that algorithm uses the relationship between enterprise and 4-nodes microstructures (quads in Figure 4), including 4-cycle, 4-chordalcycle and 4-clique for predictions.

(4) *ETF* represents using the relationship between enterprise and all microstructures (both closed triads and quads in Figure 4).

For example, we fuse information from the enterprise-enterprise link matrix $\mathbf{EE}$ and enterprise-patterns matrix $\mathbf{EP}$ for probabilistic matrix factorization(PMF) and to predict links between enterprises. In PMF, the user-item matrix $\mathbf{R} \in R^{M \times N}$ will be factorized into matrix $\mathbf{U} \in R^{M \times D}$ and matrix $\mathbf{V} \in R^{D \times N}$. $\mathbf{U}$ and $\mathbf{V}$ are latent user and item feature matrices. We use matrix $\mathbf{EE}$ and matrix $\mathbf{EP}$ to form matrix $\mathbf{R}$ shown in Figure 10. These matrices that contain links between enterprises and these microstructure

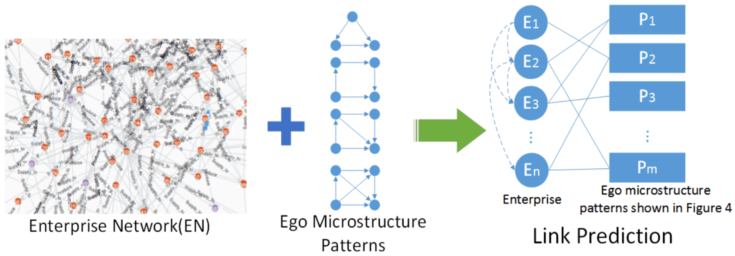


Fig. 9: *Link prediction by using these microstructure patterns.*

patterns are shown in Figure 9. We try to add the matrix $\mathbf{EP}$ to the matrix $\mathbf{EE}$ to increase the accuracy of the prediction. We let $E_{1..n}$ be the ID of enterprises in EN and $P_{1..m}$ be the microstructure patterns shown in Figure 4. The relationship between enterprises and microstructure patterns will be represented in matrix $\mathbf{EP}$. If we only use triad as feature to help prediction(*ET*), the value of m is 4 and $P_{1..4}$ are No.1-No.4 of triad shown in Figure 4(a). Similarly, when using *EF* and *ETF*, the value of m are 58 and 62 respectively. Here, k_{ij} which value is 0 or 1 represents whether there is a relationship between enterprise i and enterprise j. Meanwhile, c_{ij} which value is 0 or 1 represents whether enterprise i exists in graphlets j or not.

$$
\begin{array}{cc}
 & \begin{array}{cccc|ccccc} & \mathrm{E} & & & & & \mathrm{P} & & \\ E_1 & E_2 & \cdots & E_n & P_1 & P_2 & P_3 & \cdots & P_m \end{array} \\
\mathrm{E} \begin{array}{c} E_1 \\ E_2 \\ \cdots \\ E_n \end{array} & \left[\begin{array}{cccc|ccccc}
k_{11} & k_{12} & \cdots & k_{1n} & c_{11} & c_{12} & c_{13} & \cdots & c_{1m} \\
k_{21} & k_{22} & \cdots & k_{2n} & c_{21} & c_{22} & c_{23} & \cdots & c_{2m} \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
k_{n1} & k_{n2} & \cdots & k_{nn} & c_{n1} & c_{n2} & c_{n3} & \cdots & c_{nm}
\end{array}\right]
\end{array}
$$

Fig. 10: *Matrix*

Table 1: *Predictive Performance*

ALGORITHM	SIMRANK	PMF	GPLVM
EE	0.7579	0.8283	0.8150
ET	0.7980	0.8380	0.8532
EF	0.8373	0.8573	0.8934
ETF	0.8434	0.8575	0.9047

5.3 Experiments

Dataset. To obtain the information of Enterprise Network(EN), we we produce a dataset by crawling several open websites, including the Autohome service [1] and an open enterprise network website [2]. The data is stored in Neo4j graph database. Then we use algorithms 1, 2 to record links between enterprises and those graphlets, such as triads, 4-cycle, 4-chordalcycle and 4-clique. The matrix described in Figure 10 is constructed according to these information. There are 62,368 unique enterprises and 106,877 links between them in our dataset. These links represent the supply relationship between auto suppliers and manufacturers.
Evaluation Setting. We randomly split links between enterprises in matrix EE into training and test sets with a 4 : 1 ratio, where we use the training set for training. We also use a validation set from the training set to find the optimal hyperparameters for diffierent algorithms with different patterns. Another part EP is complete in experiment. To evaluate the accuracy of link prediction on the enterprise network, we use the Area under Curve (AUC) measure, which is considered robust in the presence of imbalance [21]. Higher AUC value indicates a better performance.

5.4 Result

The results of different algorithms are shown in Table 1. It can be seen that the results of prediction achieve improvements when using the characteristics of microstructure patterns. In using different microstructure patterns, we got different AUC rates reported in Table 1. As shown in Table 1, *ETFs* are more effective than other patterns (*EE, ETs, EFs*) for link prediction. The baseline method algorithm with *EE*, did not achieve good performance because it is difficult to predict the link between enterprises without any auxiliary information. These microstructure patterns (triad, 4-cycle, 4-chordalcycle and 4-clique) are the basic forms of the network, and can reflect the underlying structural information of EN. The results of AUC reported in Table 1 show *ETF* $>$*EF*$>$*ET*$>$*EE*. For the similar reason, 4-nodes microstructure(or graphlet) have more useful information in these methods than 3-nodes microstructure. Because we only consider closed triad in 3-nodes microstructure and 4-nodes microstructure contains more various relationship between one nodes and other three nodes. And using closed triads and 4-nodes microstructure together (*ETF*) get highest AUC rate in all the three algorithms.

[1] `http://www.autohome.com.cn`
[2] `http://www.chinaautosupplier.com`

In addtion, the microstructure patterns of corresponding enterprise will change after the prediction. For example, if predicting hypotenuse existing in 4-cycle, then a 4-cycle will become a 4-chordalcycle. Similarly, these microstructures patterns will make evolution after the link prediction.

6 Conclusion

In this paper, we proposed a fast and efficient algorithm for querying and recording closed quads from enterprise network by using the discovered triads. Furthermore, we classified the microstructure, including closed triad, 4-cycle, 4-chordalcycle and 4-clique, according to the ego network of node A. Then a statistical study on each enterprise leads us to find some interesting and meaningful rules in enterprise network. Finally, we use different state-of-the-art algorithms combining with these microstructure patterns to predict potential links on the real automobile supplying network dataset. The experimental results demonstrated the effectiveness of these microstructure patterns for link prediction.

In our future work, we plan to utilize our discoveries in predicting the possible neighbors of nodes in network, which would be well applied on recommendation systems. Furthermore, as our statistical results shown, some enterprises distinguish themselves from others on micro-structure distribution. So we would check and research these enterprises deeply.

7 ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by the National Natural Science Foundation of China (61402263, 91546203), the National Key Research and Development Program of China (2016YFB0201405), the Fundamental Research Funds of Shandong University (2016JC011), the Natural Science Foundation of Shandong Province (ZR2014FQ031), the Shandong Provincial Science and Technology Development Program (2016GGX101008, 2016ZDJS01A09), the special funds of Taishan scholar construction project and the TaiShan Industrial Experts Programme of Shandong Province No.tscy20150305.

References

1. Social structure, https://en.wikipedia.org/wiki/Social_structure
2. Dyad, https://en.wikipedia.org/wiki/Dyad_(sociology)
3. Ego, https://en.wikipedia.org/wiki/Ego
4. Ahmed, N.K., Neville, J., Rossi, R.A., Duffield, N.G., Willke, T.L.: Graphlet decomposition: framework, algorithms, and applications. Knowledge & Information Systems pp. 1–34 (2016)
5. Graphlets, https://en.wikipedia.org/wiki/Graphlets
6. Bhuiyan, M.A., Rahman, M., Rahman, M., Al Hasan, M.: Guise: Uniform sampling of graphlets for large graph analysis. In: 2012 IEEE 12th International Conference on Data Mining. pp. 91–100. IEEE (2012)

7. Biswas, A., Biswas, B.: Investigating community structure in perspective of ego network. Expert Systems with Applications 42(20), 6913–6934 (2015)
8. Dunbar, R., Arnaboldi, V., Conti, M., Passarella, A.: The structure of online social networks mirrors those in the offline world. Social Networks 43, 39–47 (2015)
9. Dong, Y., Tang, J., Wu, S., Tian, J., Chawla, N.V., Rao, J., Cao, H.: Link prediction and recommendation across heterogeneous social networks. In: 2012 IEEE 12th International Conference on Data Mining. pp. 181–190. IEEE (2012)
10. Girard, Y., Hett, F., Schunk, D.: How individual characteristics shape the structure of social networks. Journal of Economic Behavior & Organization 115, 197–216 (2015)
11. Gordon, I.R., McCann, P.: Industrial clusters: complexes, agglomeration and/or social networks? Urban studies 37(3), 513–532 (2000)
12. Huang, H., Tang, J., Wu, S., Liu, L., et al.: Mining triadic closure patterns in social networks. In: Proceedings of the 23rd International Conference on World Wide Web. pp. 499–504. ACM (2014)
13. Li, S., Daie, P.: Configuration of assembly supply chain using hierarchical cluster analysis. Procedia CIRP 17, 622–627 (2014)
14. Lou, T., Tang, J., Hopcroft, J., Fang, Z., Ding, X.: Learning to predict reciprocity and triadic closure in social networks. ACM Transactions on Knowledge Discovery from Data (TKDD) 7(2), 5 (2013)
15. Madhavan, R., Gnyawali, D.R., He, J.: Two's company, three's a crowd? triads in cooperative-competitive networks. Academy of Management Journal 47(6), 918–927 (2004)
16. Toral, S., Martínez-Torres, M.d.R., Barrero, F.: Analysis of virtual communities supporting oss projects using social network analysis. Information and Software Technology 52(3), 296–303 (2010)
17. Trpevski, I., Dimitrova, T., Boshkovski, T., Kocarev, L.: Graphlet characteristics in directed networks. arXiv preprint arXiv:1603.05843 (2016)
18. Wang, L., Liu, S., Pan, L., Wu, L., Meng, X.: Enterprise relationship network: Build foundation for social business. In: 2014 IEEE International Congress on Big Data. pp. 347–354. IEEE (2014)
19. Wasserman, S., Pattison, P.: Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp. Psychometrika 61(3), 401–425 (1996)
20. Yanardag, P., Vishwanathan, S.: Deep graph kernels. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1365–1374. ACM (2015)
21. Stager, M., Lukowicz, P., Troster, G.: Dealing with class skew in context recognition. In: 26th IEEE International Conference on Distributed Computing Systems Workshops (ICDCSW'06). pp. 58–58. IEEE (2006)
22. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: NIPS. vol. 20, pp. 1–8 (2011)
23. Lawrence, N.D., Urtasun, R.: Non-linear matrix factorization with gaussian processes. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 601–608. ACM (2009)
24. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 538–543. ACM (2002)