Talent Circle Detection in Job Transition Networks

Huang Xu†, Zhiwen Yu†, Jingyuan Yang‡, Hui Xiong‡, Hengshu Zhu‡
†Northwestern Polytechnical University, Xi’an 710072, PR China
‡Rutgers University, ‡Baidu Research-Big Data Lab
xuhuang@mail.nwpu.edu.cn, zhiwenyu@nwpu.edu.cn,
{jingyyan,hxiong}@rutgers.edu, zhiwenhengshu@baidu.com

ABSTRACT
With the high mobility of talent, it becomes critical for the recruitment team to find the right talent from the right source in an efficient manner. The prevalence of Online Professional Networks (OPNs), such as LinkedIn, enables the new paradigm for talent recruitment and job search. However, the dynamic and complex nature of such talent information imposes significant challenges to identify prospective talent sources from large-scale professional networks. Therefore, in this paper, we propose to create a job transition network where vertices stand for organizations and a directed edge represents the talent flow between two organizations for a time period. By analyzing this job transition network, it is able to extract talent circles in a way such that every circle includes the organizations with similar talent exchange patterns. Then, the characteristics of these talent circles can be used for talent recruitment and job search. To this end, we develop a talent circle detection model and design the corresponding learning method by maximizing the Normalized Discounted Cumulative Gain (NDCG) of inferred probability for the edge existence based on edge weights. Then, the identified circles will be labeled by the representative organizations as well as keywords in job descriptions. Moreover, based on these identified circles, we develop a talent exchange prediction method for talent recommendation. Finally, we have performed extensive experiments on real-world data. The results show that, our method can achieve much higher modularity when comparing to the benchmark approaches, as well as high precision and recall for talent exchange prediction.

CCS Concepts
•Information systems → Clustering;

Keywords
People Analytics; Talent Circle Detection.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. KDD '16, August 13-17, 2016, San Francisco, CA, USA © 2016 ACM. ISBN 978-1-4503-4232-2/16/08. . . $15.00 DOI: http://dx.doi.org/10.1145/2939672.2939732

1. INTRODUCTION
When there is a scarcity of skilled talents, an emerging challenge for human resource management (HRM) is how to identify the right talent from the right source in an efficient manner. Professional recruiters usually invest a lot of resources in talent source acquisition, which is focused on identifying, assessing, and engaging the sources of skilled talent candidates through proactive recruiting techniques.

As a proactive recruiting strategy, the use of Online Professional Networks (OPNs), such as LinkedIn1, for talent recruitment becomes popular in many firms. Indeed, there is rich information about talent career trajectories and the talent skills in OPNs data [17], which enables the new paradigm for talent recruitment and job search.

However, the dynamic and complex nature of this talent information imposes significant challenges to identify prospective talent sources from large-scale OPNs. More specifically, there are three unique challenges. First, there are too many people available in the network. For example, LinkedIn has reached over 400 million members around the world in October 2015. Thus, it is essential to provide a method for finding the candidate scope from many possible talent pools. Second, for different types of positions, hiring specialists usually need to consider different talent sources. Third, since people’s job transition trajectories vary a lot [16] and lack of regularity, it is necessary to investigate people’s job transition trajectories at the organizational level. Then, it is possible to capture the hidden recruitment patterns and identify the right talent sources.

Figure 1: Circles in job transition network.

To address these challenges, in this paper, we propose to use an organization-level job transition network [3], which is generated from people’s job transition trajectories. In the network, vertices stand for organizations and edges represent

1https://www.linkedin.com
the job transition amount among organizations for a time period. Also, the direction of edge indicates the job transition orientation and the weight of edge shows the quantity and the category of employees transferring between organizations. Moreover, if a given organization is used as a center node, an *ego network* can be formed by selecting all the neighbors and edges between them from job transition networks, as shown in Figure 1.

By analyzing this job transition network, we define a talent circle on the ego network to help identify prospective talent sources. A talent circle consists of organizations that have similar talent exchange patterns. With talent circles, the source of different types of talents are gathered together and can then be used for recruitment and job search. Specifically, recruiters can find candidates from organizations in most relevant circles. On the other hand, job seekers can locate the position targets in their related circles. For example, in Figure 1, the neighbor organizations of the ego have been separated into several circles with overlapping nodes or even in hierarchy. These different circles show different talent exchange patterns for talents on specific types of positions, such as financial risk analysts or software engineers.

Although the problem of ego network detection has been studied in social media [9] or co-authorship network fields [2], job transition network differs from the aforementioned scenarios in several aspects. First, the ego job transition network is much more densely connected and weighted. Since most of the organizations in an ego network have strong connections, the weight is important to distinguish the significance of edges. Second, the job transition network is direction sensitive, because the direction of edges has different meanings in different recruitment scenarios.

In light of the above, we propose a talent circle detection method for finding talent circles in the job transition networks. Specifically, we define node similarity based on talent exchange patterns and infer edge existence probability with hypothetical circles. Then, we compare the inferred probability with edge weight and refine the circle segmentation by maximizing Normalized Discounted Cumulative Gain (NDCG). Next, we label detected circles with representative words from job description to provide the semantic meaning for identified circles. Moreover, based on detected circles, we develop a talent exchange prediction method to show the effectiveness of the proposed model. Finally, we have performed extensive experiments on a large amount of real-world data. The results show that, our method can achieve much higher modularity when comparing to the benchmark approaches, as well as high precision and recall for talent exchange prediction.

2. FRAMEWORK OVERVIEW

To address the circle detection problem in job transition network, we design a framework that consists of data crawling, data transforming, modeling, learning and circle labeling. As shown in Figure 2, the framework consists of three main stages, namely network formation, feature extraction and circle detection.

**Network Formation.** In this stage, we obtain the raw data and transfer them into a formalized job transition network. The raw data are job experience records, in which each item contains a job title, a corresponding organization, a brief text description of work contents, and the beginning/end date of the job position. Specifically, we first crawl professional profiles from OPNs and transfer the inner resume into job transition trajectories by joining successive working experience items. Then, we aggregate all the trajectories into a weighted and directed network [21] at the organizational level. Finally, the ego network is defined as a subnetwork that consists of the neighbors of a specific node.

![Figure 2: The framework of talent circle detection.](image)

**Feature Extraction.** In this stage, we define similarity features according to the job transitions of nodes and the similarities between nodes. Specifically, we classify job titles into categories based on the job description posted by corresponding employees. Intuitively, organizations that exchange more employees and have more common preferences in job type distribution should be more similar to each other. Furthermore, organizations have homogeneous interactions with the ego node should be more similar than heterogeneous ones. Thus, the features used in similarity measurement can be derived from two aspects: aggregated and ego-centric personnel exchange. Specifically, aggregated personnel exchange features consist of job transition patterns centered by the node from all its neighbors; and ego-centric features consist of job transition patterns between nodes and the ego node. These features can capture the hiring characteristics of the organization. Here, the similarity is defined as the multiplicative inverse of Euclidean distance between nodes.

**Circle Detection.** In this stage, we conduct model learning process to determine node segmentation. In the model, we first put each node into one hypothetical circle and calculate the node-circle belonging coefficient based on the node similarity. Then, we remove a node from circles which it belongs to and then add it into other circles according to the level of *belongingness* and circle size. After that, we infer the possibility of edge existence based on the assumption that, if two nodes with high similarity and appear in tightly connected circles, an edge will be more possibly generated between them. Finally, we compare the inferred probability with edge weight by a customized version of NDCG and determine whether to accept the segmentation or not. The learning process is initiated as a single node circle sets and iterates until there is no update of objective function. After circle detection, we label the circles with top related organizations and keywords from job description. In particular, the learned circles can be further used for recruitment related applications, such as talent exchange prediction.

3. JOB TRANSITION NETWORK

In this section, we describe how to build the weighted job transition network, and how to extract features for measuring the similarity between nodes.
3.1 Preliminaries

We use a weighted and directed graph \( G = \langle V, E, W \rangle \) to model job transitions between organizations for a time period (e.g., in 12 months). Specifically, each node \( v_i \in V (i = 1, 2, \ldots, N) \) represents an organization, which could be a company, a university or a government department, etc. A directed edge \( e_{i,j} \in E \) from \( v_i \) to \( v_j \) stands for the aggregated job transition from organization \( v_i \) to \( v_j \). Moreover, the indegree (or outdegree) \( \text{deg}^{-}(v_i) \) (or \( \text{deg}^{+}(v_i) \)) of node \( v_i \) is the number of edges ended at (or started from) \( v_i \).

![Directed and weighted ego network.](image)

**Figure 3: Directed and weighted ego network.**

**Application-Oriented Node Degree.** In recruitment, different application scenarios concern different edge directions. For example, if the task is to discover where to find out candidates for a specific position in a company, the historical incoming transitions of that company are appropriate references. However, when job-seekers urge to know where to find the next job, they usually need to check the out-going transitions from the current company they are working for, and find out some companies which are ready for recruitment of qualified positions. Therefore, we use different degree definitions in these two scenarios. Specifically, we use \( \text{deg}(v_i) \) to represent appropriate node degree while developing model.

In the application of recruiting candidates, it is more meaningful to locate the source where employees come from, such that we set \( \text{deg}(v_i) = \text{deg}^{-}(v_i) \) where the edge weight and transfer amount are calculated based on the selected direction of transitions. Accordingly, we set \( \text{deg}(v_i) = \text{deg}^{+}(v_i) \) in the application of job position locating. As a result, the model is general for satisfying both application scenarios when using corresponding degree definition.

**Edge Weight Definition.** For a node \( v_i \) and its neighbors \( \{v_j, j = 1, 2, \ldots, \text{deg}(v_i)\} \), weight \( w_{i,j} \in W \) of edge \( e_{i,j} \) is defined as the job transition percentage between \( v_i \) and \( v_j \) within a given time period. More precisely, we use \( n_{i,c} \) to denote the total number of transitions from (or to, depends on the application scenario) a node. The edge weight is then calculated as

\[
 w_{i,j} = \frac{\tau_{i,j}}{n_{i,c}}.
\]

where \( \tau_{i,j} \) is the amount of transitions between \( v_i \) and \( v_j \). Note that, since we are only interested in one of the directions in our model for each scenario, edges will not be duplicated used in weight calculation. Furthermore, because the nodes in job transition network are densely connected, the weight is a critical indicator to distinguish the significance of edges. Thus we take the edge weight as an important determinant in both node similarity measurement and circle detection process.

**Ego Network.** The ego network in talent circle detection is a subset of the job transition network. Specifically, given a node \( v \) as the center node, the ego network is constituted of all the neighbors of \( v \). For example, as shown in Figure 3, the neighbor node set (i.e., black nodes) in ellipses forms the ego network. In particular, different application scenarios require different edge sets. The left side of Figure 3 shows a network that is suitable for identifying talent source of the ego node, and the right side network is appropriate for job-seekers to reduce job search scope when chasing a new job. Note that, because the transition patterns should not be limited in the ego network, edge weights and transitions in the ego network are measured based on the whole network \( G \), but not measured by the traffic of subset nodes. Moreover, \( u \) is not included in its ego network. We use notation \( G \subseteq G \) to denote the ego network in our model.

3.2 Node Similarity

Intuitively, the similarity of two organizations could be derived from profile information, such as whether they are belonging to the same sector, located in the same city, or whether they have common business scope.

However, static profile based similarity is not suitable in talent circle detection from the perspective of recruitment. Ideally, the similarity should reflect employees’ transition characteristics between organizations. In other words, organizations that share more job transitions and have more common preferences in job category distribution should be more similar to each other. Furthermore, since the primary goal is to find out the circles of the ego, organizations which have homogeneous interaction with ego should be more similar than heterogeneous ones. Thus, in this paper, the similarity is defined based on the historical job transitions. Specifically, the features used for similarity measurement can be derived from two aspects, namely aggregated and ego-centric personnel exchange.

**Aggregated Personnel Exchange.** This is defined as the categorical distribution of transitions to a node. In this paper, job positions in transitions are divided into 10 categories according to job title and the corresponding job description (the details are discussed in Section 6.1). The total amount of job transition \( n_{i,c} \) is separated into \( \{n_{i,c}, c = 1, 2, \ldots, 10\} \), where \( n_{i,c} \) is the amount of transitions in job category \( c \). Then, the transition proportion of each job title category with respect to the total transitions to organization \( v_i \) forms a categorical distribution vector

\[
 \alpha_i = [\alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,10}], \quad \alpha_{i,c} = \frac{n_{i,c}}{n_{i,c}}.
\]

The vector captures the constitution of employees attracted (or offered in the application of recruitment) by the organization. In Figure 4, the solid lines illustrate the source of aggregated features of in-coming traffics. From the perspective of recruitment, two organizations with similar \( \alpha \) may have analogous staffing strategies. Since the edge and the corresponding weight stand for the fraction of transitions between nodes, aggregated exchange could be treated as the extension of all the edges. Moreover, because the weight of edge is used as the objective in our model, it will be ignored in similarity measurement.

**Ego-centric Personnel Exchange.** This is the categorical distribution of transitions interfered by the ego node. From the viewpoint of the ego node, organizations may act as similar roles in talent exchange flows, even though they are different in staffing strategies. Specifically, two organizations are similar when they offering (or attracting) the same type of talents (e.g. hardware engineer) to (or from) the ego. Although the center node is excluded in its ego network, the
relationship between nodes and the ego is a determinative factor in circle detection. Particularly, two nodes are more likely to co-appear in a circle when they have common interaction patterns with the ego node. In fact, two organizations may be grouped into the same circle because they feed the ego with the same category of employees, even though they do not have the same organization scale or in the same sector. Ego-centric features are defined based on $\alpha_i$, while the transitions between node $u$ and $v_i$ are considered. We use $\mu_i$ to denote the categorical distribution vector of transitions between $u$ and $v_i$, and then define the feature vector as

$$\beta_i = [\beta_i^1, \beta_i^2, ..., \beta_i^{10}], \beta_i^k = \frac{\mu_i^k}{\alpha_i^k} \cdot w_{u,i},$$

(3)

where $\alpha_i$ is the aggregated feature and $w_{u,i}$ is the weight of edge between $u$ and $v_i$. Each dotted line in Figure 4 illustrates a dimension of ego-centric features of node $v_i$ and $v_j$, the ego. $\beta_i$ is the fine-grained transition categorical distribution normalized by aggregated transition categorical distribution. Indeed, two similar vectors indicate they interact with the ego in similar way, since the vector encodes the interaction pattern between the ego and organization. Based on the feature vectors $\alpha_i$ and $\beta_i$, similarity $\text{Sim}(v_i, v_j)$ is defined on the merged vector $F_i = [\alpha_i, \beta_i]$ and $F_j = [\alpha_j, \beta_j]$, as shown in Equation 4.

$$\text{Sim}(v_i, v_j) = \frac{1}{||F_i - F_j||_2},$$

(4)

4. CIRCLE DETECTION MODEL

In this section, we introduce the details of our model for talent circle detection in the ego job transition network.

4.1 Basic Concepts

In this paper, we design a generative model to detect overlapping and hierarchical talent circles. As in the ego job transition network, two organizations should be assigned to a circle if they are closely connected and also similar to each other. The basic assumptions of our model have three aspects. First, two nodes have a possibility to form an edge when they are co-appearing in a circle. Second, circles that contain stronger connected nodes lead to a higher possibility to form edges among contained nodes. Third, two nodes in a circle with higher belongingness also have a higher edge existence possibility.

The probability of edges is generated based on the above assumptions. We then compare the inferred probability with the edge weight in an ordered manner. The objective of the model is to maximize the weighted similarity between the weight order of neighbors and the inferred probability of edges around an organization.

To build up the metrics of our model, we first define the concepts of natural ordered neighbor sequence, circle and belonging coefficient as follows.

**Natural Ordered Neighbor Sequence.** Given a node $v_i$, the natural ordered neighbor sequence $A_i$ is the descending ordered list of neighbors according to the edge weight $\{w_{i,k}| e_{i,k} \in E\}$. Formally, we have

$$A_i = (v_1, v_2, v_3, ..., v_j, v_{j+1}, ...),$$

where $w_{i,j} \geq w_{i,j+1}$. We use $A_{i,k}$ to refer to the $k$-th node in $A_i$. For example, if $v_i$ connects to $(v_1, v_2, v_3)$ and weights among them satisfy $(w_{1,2} > w_{1,1}, w_{1,1} > w_{3,3})$, then $A_i = (v_2, v_1, v_3)$. The neighbors with the same edge weight are arranged by the indexing (e.g., alphabetic) order. When constructing the model for a specific organization, the ego node is excluded from the neighbor list.

**Talent Circle.** A circle is a subset of neighbors of an ego node, where nodes within a circle are closely connected and similar to each other. We use $\{C_m \in C\}$ to denote circles, where $m = 1, 2, ..., M$ and $C_m \subseteq V$. Intuitively, on the one hand, if the circles are appropriately detected, nodes with strong connections should more likely be settled in the same circle than the nodes with weak connections. On the other hand, from the perspective of the node similarity, similar nodes are more likely to appear in the same circle. Meanwhile, circles could be hierarchical, which means circles with strongly connected nodes could be contained in circles with weakly connected nodes. Moreover, circles could also be overlapping. In other words, a node can belong to more than one circle.

**Belonging Coefficient.** Since circles may be overlapping, a node could more tightly belong to a circle than another circle. The belonging coefficient is defined as the tightness strength of a node to a circle. Specifically, we use a belonging matrix $U_{i,m} \in \mathbb{R}(0 \leq U_{i,m} \leq 1, i \in 1, 2, ..., N, m \in 1, 2, ..., M)$ to denote the belonging strength of node $v_i$ to circle $C_m$, and a larger value of $U_{i,m}$ to indicate a tighter belongingness. In our model, $U_{i,m}$ is defined as the average mutual similarity between node $v_i$ and the rest of the nodes in circle $C_m$, as shown in Equation 5. When there is only one node in a circle, the belonging coefficient is assigned to 1. In particular, the belonging coefficient affects the inferring of the edge existence probability in the model as well as the node-circle belonging dynamics in the learning process.

$$U_{i,m} = \frac{1}{|C_m| - 1} \sum_{v_j \in C_m, j \neq i} \text{Sim}(v_i, v_j).$$

(5)

4.2 Model Formalization

According to our assumptions, a pair of nodes within a circle has a possibility to generate an edge, and the generated edge weight is determined by two factors: the number of common circles they belong to, and the belonging coefficient of the node pair to the common circles. In reality, if two companies exchange employees more frequently, they should be more likely to be placed in the common circles. We use an iterative method to approximate the real segmentation and start from a simple temporal status where circles are hypothetically existing. Accordingly, the probability of two nodes forming an edge is defined based on node-circle belonging relation in the temporal status.

In a temporary status, all nodes are assigned into circles. We define an indicator $\rho(i,j)$ to capture the edge
existence possibility (or weight strength) between node \(v_i\) and \(v_j\). Specifically, the value of \(\rho(i,j)\) is high if \(v_i\) and \(v_j\) are put into a common circle and have high belonging coefficients to the circle. For each node \(v_j\) which satisfies \(|\{C_m|\{v_i, v_j\} \subseteq C_m\}| > 0, \rho(i,j)\) is defined as follows:

\[
\rho(i,j) = \exp \left\{ \sum_{\{C_m|\{v_i, v_j\} \subseteq C_m\}} \frac{C_{i,m} \cdot C_{j,m}}{\text{sim}(i,j)^{-1} - T_m + \eta} \right\}, \tag{6}
\]

where \(T_m\) is the minimal \(U_{j,m}\) for all the \(v_j \in C_m, \) \(\eta\) is the maximal \(T_m\) for all circles \(C_m). In other words, \(T_m\) is a similarity threshold of circle \(C_m), which indicates the tightness of nodes within a circle. \(\eta\) is used to guarantee that the value of \((-T_m+\eta)\) is positive and has a higher value when \(T_m\) gets smaller. A smaller value of \(T_m\) reduces the value of \(\rho(i,j),\) such that a pair of nodes in lower similarity threshold circles generates lower \(\rho(i,j)\) value. On the contrary, \(U_{i,m}\) and \(U_{j,m}\) guarantee that \(\rho(i,j)\) has a high value if both of the belonging coefficients are high.

Similar to the natural ordered neighbor sequence \(A_i), we arrange all the neighbors of node \(v_j\) according to \(\rho(i,j)\) in descending order as \(B_i = (..., v_j, v_{j+1}, ...), \) where \(\rho_{i,j} \geq \rho_{i,j+1} \).

Let us use \(B_{i,k}\) to refer to \(k\)-th node in \(B_i. \) Ideally, given a pair of nodes \((v_j, v_i)\), if \(w_{i,j} > w_{i,l}\) and the edge weight is well inferred, it should have \(\rho_{i,j} > \rho_{i,l}. \) The opposite relation indicates wrongly assigned circles. Thus we treat the sequence difference between \(A_i\) and \(B_i\) as the goodness of circle division. If \(w_{i,j} > w_{i,l}\) and \(\rho_{i,j} < \rho_{i,l},\) then the pair of nodes \((v_j, v_i)\) is an inversion of \(A_i\) and \(B_i. \) The inversion number is the total inversions between two sequences as \(\text{inv}(v_j) = |\{v_j, v_i| w_{i,j} > w_{i,l} \cap \rho_{i,j} < \rho_{i,l}\}|. \) According to the definition, there is no inversion when edge weights are perfectly inferred. In addition, the more error exists, the larger inversion number will be conducted.

However, the calculation of inversion number is computational complex, because the length and elements of \(A_i\) and \(B_i\) are not always match. Fortunately, since what we need is a metric that can measure the difference of two weighted sequences, we propose to use Normalized Discounted Cumulative Gain (NDCG)\([7]\) for measuring the differences instead. Specifically, NDCG is a measure of ranking quality in information retrieval by calculating the weighted ranking results according to ideal ranking lists. When replacing document with neighbors, and relevance with edge weight in the definition, NDCG is suitable for measuring the difference between \(B_i\) and \(A_i. \) The basic assumptions of NDCG can be grafted on to measure sequence difference smoothly:

- Highly weighted neighbors are more meaningful when appearing earlier in \(B_i.\)
- Highly weighted neighbors are more meaningful than marginally weighted neighbors, which are in turn more meaningful than unconnected nodes.

By the migration of conceptions, Discounted Cumulative Gain (DCG) of node \(v_i\) is defined as follows.

\[
\text{DCG}_i = \sum_{k=1}^{\text{deg}(v_i)} \frac{w_{i,k}}{\log_2(k+1)}, \tag{7}
\]

where \(k\) is the sequence order index in \(B_i, \). According to the definition, \(\text{DCG}_i\) increases when a node with higher weighted edge be moved before a lower one in \(B_i. \) In other words, \(\text{DCG}_i\) is high if \(B_i\) is close to \(A_i, \) especially when the prefix of \(B_i\) is close to the prefix of \(A_i. \) Ideal \(D\text{CG}_i\) (\(IDCG_i\)) is calculated based on the neighbor sequence of \(A_{i,k} \) and it is the upper bound of \(D\text{CG}_i\).

\(N\text{DCG}\) is then defined as normalizing \(D\text{CG}_i\) by \(IDCG_i. \) Note that \(B_i\) is the list of neighbors that co-appear with \(v_i\) in circles, so the length of \(B_i\) is usually not equal to \(|A_i|\) in reality. Typically, \(B_i\) may contain many elements as suffix that does not belongs to \(A_i. \) For example, if nodes in \(A_i\) are \((v_1, v_2),\) then \(B_i = (v_1, v_2, v_3, v_4)\) and \(B_i' = (v_1, v_2)\) will have the same \(N\text{DCG}_i. \) Thus we use set difference \(A_i \triangle B_i\) to penalize it, and define objective function of our model in Equation 8:

\[
F_U = \sum_{i=1}^{N} \left( n_i \cdot \frac{\text{DCG}_i}{\text{IDCG}_i} \cdot (1 - \frac{|A_i \triangle B_i|}{|A_i \cup B_i|}) \right), \tag{8}
\]

where \(U\) is the belonging coefficient matrix and acts as a dynamic parameter in the learning process. Meanwhile, \(F_U\) is related to the proportion of transitions such that a larger node has a higher priority.

\section{Model Learning Algorithm}

In this paper, we propose to use a bottom up learning algorithm starting from a simple node-circle belonging status to maximize objective Equation 8, as shown in Algorithm 1.

In the initial step, each node \(v_i\) is assigned to a corresponding circle \(C_i,\) such that \(C_i = \{v_i\}, i = 1, 2, ..., N, \) \(U\) is an identity matrix, since the node-circle belonging coefficient is 1 if there is only one node in the circle. For each node, we first compute the natural ordered neighbor sequence \(\{A_i\}\) according to \(\{w_{i,k}|v_i, k \in E\}\). As illustrated in the left side of Figure 5, at step 0, \(U^0\) is an identity matrix that encodes initial node-circle belonging relationships.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.png}
\caption{Learning process illustration.}
\end{figure}

In the learning process, the algorithm iteratively remove a node from circles it belongs to and add the node into other circles to maximize the objective function in Equation 8. In each step, the leaving and joining circles are determined by both the belonging coefficient and the circle size. Let us use notation \(NC_i^t = \{C_m|v_i \in C_m\}\) to denote node-circle membership in iteration \(t, \) which is a binary version of non-zero elements in the corresponding row of \(U_{i,m}.\) The learning process then changes \(NC_i^t\) by adding and removing nodes in circles, and evaluates the value of \(F_U\) of the temporal status to decide whether accept or reject the change.

Specifically, in iteration \(t,\) for each node \(v_i,\) we first calculate the edge existence possibility according to Equation 6,
Algorithm 1 Talent circle detection by maximizing NDCG

Require: Similarity matrix $SIM, IDC\!G, A_i$, Maximal iteration steps $\varphi$ 
Ensure: Node-circle belonging coefficient matrix $U$
Initial: $U_{i,k} \leftarrow I, NC_i^0 \leftarrow \{v_i\}, F \leftarrow 0$
while $F$ changed in last $\varphi$ steps do
    Calculate $\rho_i^t$, then Arrange $B_i$
    for Node $v_i \in V$ do
        $d_{rm} \leftarrow |NC_i^0| \cdot |B_i|/|A_i|$
        $r_{mc} \leftarrow$ sample $d_{rm}$ circles from $NC_i$ by $P_r(i,m)$
        $NC_i^t \leftarrow$ setdiff($NC_i^0, r_{mc}$)
        $d_{add} \leftarrow 1/|NC_i^t| \sum_{i} |NC_i^t - |NC_i^0| \cdot |A_i|/|A_i|$
        $adc \leftarrow$ sample $d_{add}$ circles by $P_a(i)$
        $NC_i^t \leftarrow$ setunion($NC_i^t, adc$)
    end for
    Calculate $U_{tj}$
    $F_U \leftarrow F_U(NC_i^t, B_i, SIM, IDC\!G)$
    if $F_U > F$ then
        $F \leftarrow F_U$
        $NC_i \leftarrow NC_i^t$
    end if
end while

Job Transition Transformation. Items in job resume are transformed into job transitions by joining successive job record pairs. Specifically, we sort the resume items by starting time in ascending order, and then compare the end date to identify job-hooping activities, as shown in Figure 6. If the end date of record $j$ satisfies $et_j < et_{j+1}$, then we treat home” are neglected. Table 1 shows the detailed statistics of our data set.
There are three steps of data pre-processing, including job transitions transformation, job title categorization and transition network initialization.

<table>
<thead>
<tr>
<th>Data</th>
<th>Capacity</th>
<th>Data</th>
<th>Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resume</td>
<td>1,980,000</td>
<td>Job title</td>
<td>284,245</td>
</tr>
<tr>
<td>Transition</td>
<td>2,123,383</td>
<td>Category</td>
<td>10</td>
</tr>
<tr>
<td>Organization</td>
<td>649,076</td>
<td>Average flow</td>
<td>30</td>
</tr>
</tbody>
</table>

6. EXPERIMENTAL RESULTS
In this section, we evaluate our talent circle detection model with extensive experiments on real-world data.

6.1 Experimental Data
In this paper, we use a representative real-world data set from one of the largest commercial OPNs, to study the circle detection problem. Specifically, the data set contains 1.38 million professional profile pages from Sept. 2014 to Dec. 2015. Each page contains a job experience list in which each item shows the job title, the organization, the time period in a monthly granularity and a brief text about the job description [17]. According to the URL of organizations in resume records, there are 649,076 organizations in total. Since we are focusing on job-hooping behaviors among organizations, work experience records with keywords indicating freelancers like “freelance”, “self employed” and “work at

Figure 6: Transforming resume into job transition.
the start date of successor and the job title is determined by the title of predecessor when it changed during a transition.

Common transition date lasts from 1970s to the date when we collect data, and most of the transitions happening after 2000s. As shown in Figure 7, job transition frequency has a clear cycle of 12 months, so we use 12-month as a time window in our experiments.

![Job Transition Frequency](image)

**Figure 7: Aggregated job transition frequency.**

**Job Title Categorization.** On professional profiles, users are free to specify their job titles, and usually write a short text description of their work [1]. In total, there are 284,245 different titles in our data set. In our experiments, we categorized the job titles into a few dozens of classes according to job functions by using online API from a third-party tool called Autocoder 2, which can classify job related content such as resumes and job description to a standardized hierarchy of occupation categories known as the Occupational Information Network. After that, we manually classified titles into 142 classes and further integrated into 10 major categories, as shown in Table 2.

**Table 2: Job title category.**

<table>
<thead>
<tr>
<th>Category</th>
<th>#Subclass</th>
<th>#Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Tech</td>
<td>11</td>
<td>54,091</td>
</tr>
<tr>
<td>Finance</td>
<td>10</td>
<td>20,195</td>
</tr>
<tr>
<td>Professional and Business Services</td>
<td>21</td>
<td>75,565</td>
</tr>
<tr>
<td>Transportation</td>
<td>6</td>
<td>7,510</td>
</tr>
<tr>
<td>Consumer Goods</td>
<td>14</td>
<td>11,548</td>
</tr>
<tr>
<td>Goods-Producing Industries</td>
<td>7</td>
<td>7,850</td>
</tr>
<tr>
<td>Public Administration</td>
<td>22</td>
<td>27,994</td>
</tr>
<tr>
<td>Leisure and Hospitality</td>
<td>19</td>
<td>32,553</td>
</tr>
<tr>
<td>Education and Health Services</td>
<td>14</td>
<td>36,612</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>19</td>
<td>10,327</td>
</tr>
</tbody>
</table>

**Transition network initialization.** We aggregated all the transitions at the organizational level and formed a job transition network. All the transitions from organization $i$ to organization $j$ form a directed edge starting from $i$ and ending at $j$ with the transition volume $n_{i,j}$. In average, there are 30 transitions between two organizations, and 84% of the flows happen among 20% of the nodes. Among all the transitions, approximately a quarter of the transitions took place within the same organization, and we ignored these loops because the ego is excluded in talent source identification.

### 6.2 Modularity based Evaluation

To evaluate the performance of our model in terms of circle detection, we compare our model with two state-of-the-art algorithms on modularity [10], which is a widely used metric in community detection [15]. Since the original definition of modularity is defined on non-overlapping communities, in our experiments we use an extension introduced in [12]. Specifically, we first choose two latest algorithms as baselines, namely Coordinate Ascent (CA) from social network analysis [9] and Simulated Annealing (SA) from co-authorship network analysis [2]. Meanwhile, we choose two classic algorithms as baselines, namely Random Walks (RW) based community detection [13] and Edge Betweenness (EB) based method [11]. Since there are some differences between network definitions, we follow the impliction of RW and EB in [5], and from original websites 3 4 of four methods, respectively.

**Table 3: Comparison results of modularity.**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Our Model</th>
<th>CA</th>
<th>SA</th>
<th>RW</th>
<th>EB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.5636</td>
<td>0.5536</td>
<td>0.5652</td>
<td>0.5464</td>
<td>0.5420</td>
</tr>
<tr>
<td>Our Model</td>
<td>0.6376</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 8: The comparison results of modularity.**

Because the belongingness is usually normalized to have a summation of 1, we normalize the belonging coefficient $U_{i,m}$ by its row sum, i.e., $\overline{U}_{i,m} = \frac{U_{i,m}}{\sum_{m} U_{i,m}}$, and we set belonging coefficient of nodes as the multiplicative inverse of belonging circles in CA and SA. The coefficient is set to 1 in RW and EB since there is no corresponding definitions in these two methods. Since there is a modularity in each ego network, we pay more attention to the performance on majority egos and comparing the corresponding results of ego in other algorithms, as shown in Figure 8.

The result is based on the experiments on 7,000 egos which are the top 10 percent nodes in degree. There is an average of 5.6 circles and 3.1 organizations in each circle. As shown in Figure 8, the average modularity are 0.6376 (Our Model), 0.5636 (CA), 0.5536 (SA), 0.5464 (RW), and 0.5420 (EB), respectively. As a result, our method outperforms all other methods at least 7.2% in average. The majority of results of our method are also higher comparing to other methods in despite of few outliers with low modularity. The results clearly indicate that our method is more accurate for circle detection in terms of modularity in the weighted and densely connected job transition networks.

### 6.3 Circle Case Analysis

According to the model assumptions, organizations within a circle should have common interactions with the center organization and also exchange similar employees among each other. Thus different circles have different job transition characteristics. To this end, we select the results of several organizations for analysis.

To interpret the results, we rank the organizations in the circle according to the belonging coefficient and select top

3. http://cse.iitkgp.ac.in/resgrp/cnerg/circle/
ones as representative organizations. Besides, we extract related keywords from job description to facilitate the interpretation. We trace job transitions between ego and the set of companies in the circle, and extract job descriptions of these transitions. All the descriptions related to a circle are joined into a document, and thus a set of documents form as a corpus. After a removal of stop words, we arranged the words in a document according term frequency inverse document frequency (TF-IDF), and manually select several representative keywords in top 10% TF-IDF words.

Figure 9: Circles of AOL. Organizations in circle (A) provide quantitative analysts. (B) provides digital marketing specialists. (C) provides customer operation support. (D) provides sales marketing analysts.

As shown in Table 3, we display the representative organizations and descriptive keywords for three ego examples, namely Citi Bank, AOL, and Accenture respectively. Four circles of each ego have been listed with corresponding keywords for job title description. For instance, since Citi is a corporate financial organization, most of the organizations in the Citi circles are focus on financial services. Although circle 1, 2, 3 of Citi are banking or related financial consulting services, the characteristics of talent sources from each circle are very different. For example, according the description of keywords, circle 1 mainly involves the talents for tasks such as corporate finance or project management, change management. While circle 2 focuses on job applicants specialized in corporate and financial risk management tasks, and circle 3 deals with talents for trading services, capital markets, or investment banking. The companies in circle 4 are mainly in IT services industry and the talent source from this circle are focusing on quantitative research, desk strategist, and quantitative analytics for Citi bank.

In summary, most of the companies in the circles for the ego are in the similar industry of the ego company, but different circle shows different needs of talent specialties. Similar observations can be found on the media and advertising company AOL (as illustrated in Figure 9) and the very successful consulting company Accenture. Keywords of job title summary based on the the top keywords of description is also listed in the last column of the result table.

6.4 Talent Exchange Prediction

One of the major application of talent circle detection in job transition network is to identify the source organization for selecting candidates in hiring. In fact, the detected circles are appropriate references for hiring different types of employees. For example, when the goal is to hire software engineer, recruiters can refer to the organizations appearing in circles which have labels in “computer software”, “software engineer” or “programmer”, etc. It is possible to precisely locate target organization by using the circles.

Figure 10: The feature distribution of typical circles of Citi bank.

However, enumerating related labels is inefficient when selecting circles for a job title category. As users are free to describe their works, there are too many alternative words for a title. There is a high possibility of mismatching by simply indexing keywords to solve the problem. Therefore, we design a quantitative method to identify related circles by ranking the average feature vectors. Specifically, the average feature of a circle $C_m$ is defined as a vector $FC_m$ where each dimension is defined as

$$FC_m^i = \frac{1}{|C_m|} \sum_{v_j \in C_m} a_j^i.$$  

For a given ego, $FC_m$ can be normalized by subtracting the average circle feature $\frac{1}{M} \sum_m FC_m$, where $M$ represents the number of circles. The normalized $FC_m$ is treated as the “feature of a circle”. Figure 10 shows the typical normalized circle feature distribution of ego “Citi”, in which each area is the value of a circle feature in 10 dimensions. It indicates that different circles have different feature peak dimension combination [20]. This phenomenon also appears in circles of other egos. It suggests that average feature value is a reasonable way to identify different circle properties. We can then rank the circle features by different dimensions and select top ranked circles as related candidates.

Talent Exchange Prediction Problem. This problem refers to predict where employees will leave for from an ego, based on the detected circles. In real job market, majority of employees will go to the same set of companies where the forerunner in the same job type went, because the sector usually keeps stable and forms some specific job hooping trajectories. Although other employees may go to other companies, it is highly possible that they still stay in organizations which have similar labor demand. Therefore, we predict that the employees in a given category will go to the circles that labeled in the same category. By taking advantage of circle features, we design a simple talent exchange prediction mechanism.

For each given job title, we predict where employees with the title will leave for. Specifically, we first rank the circle feature [22, 19] according to the corresponding dimension in a descending order, and then select companies in top circles as the predicted destination. We use the job transition destination from the ego in the next year as ground truth. The prediction and evaluation procedures have three stages.
Table 3: Top circles: representative organization and keywords.

<table>
<thead>
<tr>
<th>Ego</th>
<th>Circle</th>
<th>Top Organization</th>
<th>Top Keywords</th>
<th>Job Title Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citi Bank</td>
<td>1</td>
<td>Capco, RBS Markets and International Banking, HSBC, PwC, Lehman Brothers, HSBC Global Banking and Markets, USAA</td>
<td>project, process, management, finance, back office, retail, firmwide, mysql, query, infrastructure</td>
<td>Corporate, finance, Procurement, management, Change management</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>UBS, Credit Suisse, Barclays Investment Bank, Bank of Montreal, Newmark Grubb Knight Frank</td>
<td>stock, corporate, cash, compensation, extract, matlab, realtime, portfolio, oracle, model, data</td>
<td>Corporate/financial risk management</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>YES Bank, John Hancock Financial Services, HSBC Private Bank, CHASE</td>
<td>finance, insurance, banker, private, systematic, equity, mortgage, rate, risk, business</td>
<td>Trading and Capital markets, Investment banking</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Sun Microsystems, Google, Palantir Technologies, D+H, Intel Corporation</td>
<td>visual, linear, predict, calculate, curve, java, system, algebra, compute, diagram, variable</td>
<td>Quantitative analytics, Desk strategist</td>
</tr>
</tbody>
</table>

| AOL      | 1      | Millennial Media, popsugar, Discovery Networks Benelux, WSJ, Wunderman DC, Comcast, Townsquare Media | social, live, media, content, week, brand, advertise, broadband, freelance, independent, local, publish, editor | Digital marketing |
|          | 2      | Thomas Cook, Foursquare, Last.fm, Radio One, Wunderman | accuracy, analyze, online, metric, sale, customer, website, trend, user response, webdata, revenue, adhoc, analyze, dashboard | Sales marketing, Digital sales |
|          | 3      | Comcast, Adconion Media Group, Abril, Hearst, Los Angeles Times, Amdocs, WSJ | center, call, service, agent, maintain, quick, complaint, response, issue, communicate, customer, monitor | Customer operation, Sales operation and support |

| Accenture| 1      | cognizant, logica, deloitte digital, hewlett packard enterprise, avanade, gartner, gtech | technology, enterprise, system, application, communication, defencework, enterprise, architecture | Business analyst, ERP, CRM, Enterprise Architecture |
|          | 2      | capgemini consulting, deloitte, ey, alexander mann solutions, towers watson | sgl, management, waterfall, employee, execution, procurement, trainer | Technology project management |
|          | 3      | deloitte, morgan stanley, american express, pwc, standard chartered bank, alexander mann solutions | e-source, spend, procurement, pension, finance, pay, pipeline, workforce, resource | Financial and Accounting services |
|          | 4      | microsoft, ibm, google, oracle, amazon, elhay, cisco, facebook, tata consultancy, csc, supra steria, emc, avanade, altran | digit, advance, project, innovation, deploy, data, interface, module, architecture, technology, crm, methodology, network, database | System or Enterprise integration and implementation |

First, for a given ego and title, we rank circle features by corresponding dimension in descending manner. Then we select organizations in top ranked circles as candidates. Finally, we compare predicted organization candidates with the companies which employees in the same title category leave for in the next 12 months.

![Figure 11: The precision and recall performance of talent exchange prediction.](image)

We measure the precision and recall to evaluate the performance of talent exchange prediction. The circles are detected based on the data that have transitions during January 2013 and December 2013, and test data are the transitions during January 2014 and December 2014. Figure 11 shows the average, upper and lower bound of precision and recall for 10 job categories of Citi. As shown in the result, the precision and recall depend on the number of circles we used in prediction. We find that, the precision firstly increases with the increase of circles, and then reaches a peak of 73% in 4-5 circles, while it finally decreases when the number of circles larger than 6. The recall increases consistently with the number of circles, and stays stable after 6 circles in 65%. The results indicate that a majority of employees in a given job category will leave for companies that appears in the circles with corresponding characteristics. It suggests that the detected circles are appropriate node separation from the perspective of talent exchange flow.

7. RELATED WORK

In this section, we review two categories of literatures that are related to this paper, namely research on data mining for recruitment analysis, and research on circle detection in social networks.

Data mining for recruitment analysis. Recent years have witnessed the increasing popularity of using data mining techniques for addressing human resource management (HRM) problems [14]. The recruitment process is one of the important sub-domains of the HRM. Although many classic data mining tasks, such as classification, association rules and clustering [4, 6], have been performed to recruitment for personnel selection and talent prediction, only few existing work is focus on the analysis of job transition network for recruitment. For example, [3] proposed a real-time system for mining job-related patterns from social media by analyzing the job transition network. [17] showed a novel approach for modeling the professional similarity by mining professional career trajectories.
Circle detection in social networks. Social circle detection in ego networks is first proposed by McAuley and Leskovec [8, 9]. They proposed the node clustering problem and developed a model for identifying the circles including both network structure and user profile information from several popular social networks. Recently, [2] applied the techniques of automatic circle detection in an ego network on the field of co-authorship network by proposing an unsupervised method that combines both various node features and node similarity measures. Also, [18] used a multi-view clustering method for automatically detecting the social circles. Due to the different background in the job transition networks, although we have similar problem of discovering circles in the ego networks, the node features and node structures are very different. Therefore, different objective function and learning process have to be designed specifically for the job network scenario. To the best of our knowledge, this is the first attempt to detect circles for each company in a job transition network. We believe the findings from our model can further help to enhance the effectiveness of human resource tasks, such as staffing.

8. CONCLUSION

In this paper, we investigated how to identify the right talent sources for recruitment from Online Professional Networks (OPNs). Along this line, we first created a job transition network based on job transition trajectories at the organization level. Then, we proposed a talent circle detection model for extracting talent circles from the job transition network in a way that every circle includes the organizations with similar talent exchange patterns. With the help of these talent circles, the organizations can find the right talent for recruitment and the job seekers can locate suitable jobs for themselves. Moreover, based on these identified cycles, we developed a talent exchange prediction method to predict the possible destination companies for the job hopping employees. As shown in the experimental results on real-world OPNs data, our approach outperformed the benchmark methods in terms of modularity.

9. ACKNOWLEDGMENTS

This research was supported in part by the National Basic Research Program of China (No. 2015CB352400), the National Natural Science Foundation of China (No. 71329201, 61332005, 61402369), Microsoft, and the Rutgers National Natural Science Foundation of China (No. 71329201, 61332005, 61402369), the National Research Program of China (No. 2015CB352400), the National Science, Engineering and Technology, Penang, Malaysia, 2009.

10. REFERENCES