Exploiting Synchronicity Networks for Finding Valuables in Heterogeneous Networks

Zhen Wen [*] Ching-Yung Lin [†]

Abstract

Successful enterprises depend on high performing teams consisting of productive individuals, who can effective find valuable information. It is highly desirable yet challenging to identify these inter-related, multi-typed entities that can potentially help to improve enterprise performance, based on observation of their dynamic behavior in the organizational social networks. In this paper, we propose a novel approach to analyze and rank heterogeneous objects in multi-level networks by their value for improving productivity. Compared to existing approaches that either focus on static factors of productivity or single type of entities (*e.g.*, individuals), our work offers two unique contributions. First, we propose a novel multi-level synchronicity network representation which allows us to exploit the structural characteristics of various entities' dynamic behavior. Furthermore, based on the synchronicity networks, we propose a novel algorithm for **H**eterogeneous **M**ulti-level networks **R**anking, to simultaneously rank inter-related heterogeneous entities (*e.g.*, topics, individuals and teams) by their value. Our experiments demonstrate that our approach significantly outperforms existing methods in both enterprise organizational social networks as well as public social media.

1 Introduction

Social networking tools have provided unprecedented opportunities for individuals and organizations to interact with each other and tap into diverse information. It also brings up a challenge to identify "valuables" from explosive amount of data that can benefit people and organizations' productivity. For example, it is highly important for enterprises to identify potential top performing teams, based on their characteristics similar to past observation of high performing teams' behavior. It would help enterprises to better select teams that are likely to contribute more to the enterprise's growth. Moreover, the enterprises need to understand how the teams are composed. For instance, who are the productive individuals in the teams and how they interact with each other. This knowledge can assist the enterprises to better build effective teams for their success. In addition, it is highly desirable that the enterprises learn more about their top performers such as what types of information they tap into for expanding their horizons. This would allow the enterprises to better train and grow their employees. In a similar spirit, it is desired to find inter-related, heterogeneous entities[1] of high value on social media, such as influential communities, individuals and viral topics.

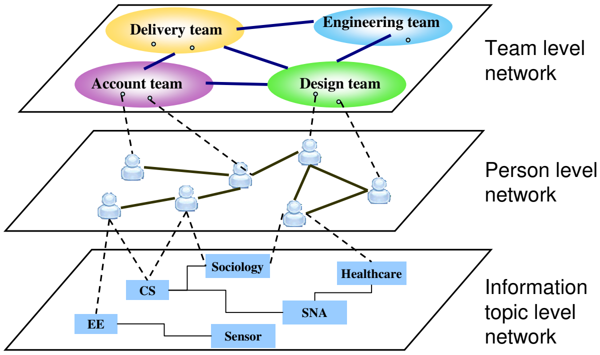


Figure 1: The multi-level composite networks of teams, individuals and information topics in enterprises.

To achieve the aforementioned goal of finding high value, heterogeneous entities in composite networks, computational models are demanded to correlate their value and observable characteristics. Structural features of the networks have been shown promising. For example, centrality measures such as PageRank [6] are effective in ranking web pages. Structure holes [7, 13] are demonstrated to be related to people's productivity. Recently, Saavedra *et. al.* [18] show that dynamic features in social networks such as synchronicity, correlate to financial traders' performance. However, it still remains a challenge to effectively find highly productive

[*] IBM T.J. Watson Research Center
[†] IBM T.J. Watson Research Center

[1] We use the terms "entity" and "object", "network" and "graph" interchangeably in this paper.

people, especially in organizational settings. For example, in enterprises one of employees' productivity metrics is their contribution to the company (*e.g.*, generated revenue), which may be related to the employees' observable behavior (*e.g.*, participation of social software) but the precise correlation is unknown. In addition, the demanded computation models need to address the challenge of ranking heterogeneous entities in composite networks where there are teams, individuals and information.

This paper presents an approach toward addressing this challenge. In particular, we focus on the dynamic aspects of people's interactions in their social networks. As synchronous behavior signals people's ability to collaborate and utilize collective wisdom, we hypothesize that high performers may selectively synchronize with other people, teams and topics. Thus, understanding the structure of synchronicity can improve finding valuables in composite networks. We construct multi-level synchronicity networks capturing the coordination among teams, individuals and information topics, respectively. The composite synchronicity networks can be utilized as a proxy to estimate how much teams and individuals may effectively engage their collaborators in actions, and how much information topics dynamically influence each other. We then propose an algorithm to rank the heterogeneous entities by value, and demonstrate its effectiveness in a rich socio-info dataset collect in a large international enterprise with more than 400,000 employees. The dataset contains various types of data sources including email, instant message communications, calendars, hierarchical organizational structure as well as performance metric in terms of revenue generated [13]. In addition, we validate the effectiveness of our approach in public social media data such as Twitter, for the task of finding users that can be high quality information sources.

To the best of our knowledge, this is the first study of computational models for ranking teams, individuals and information by productivity without having to access performance data. Instead, it is based on observations in social networks within large enterprises. This paper offers two unique contributions. First, we propose synchronicity networks, a graph-based *dynamic* feature of people's interactions, which can be utilized to effectively predict people's productivity. Second, to find heterogeneous valuable entities related to top performers (e.g., teams and information), we propose a novel algorithm to simultaneously rank heterogeneous entities in composite networks. The advantages of our approach are validated in both enterprise social networks as well as public social media.

2 Related Work

Our study on exploiting composite synchronicity networks for ranking valuable objects is related to several areas of research including prior studies on people's social networks and dynamic behavior in relation to their productivity, and previous approaches on composite network analysis that contain heterogeneous objects. We review a brief sampling of related work in each of these areas.

2.1 Social Network Analysis and Productivity

Various studies of organizational social networks have been conducted to understand the relationship between social networks and productivity. Burt [7] has shown the important influence of social network topologies on productivity. In addition, knowledge and expertise accumulated within one's social contacts can have a significant consequence on worker performance [17]. Recently, researchers [13, 23] analyzed a large scale organizational social network and found both network topology (*e.g.* structure holes) and node attributes (*e.g.*, strong ties to appropriate human capital) can be beneficial to worker performance.

Researchers have also found that the dynamic aspects of social networks are related to productivity. Synchronicity is one of these dynamic aspects which focuses on how people's dynamic behavior are aligned temporally. Saavedra *et. al.* [18] find that the synchronicity of a group financial traders' trading activities positively correlate to their trading performance. Adarov *et. al.* [1] study the network of 58 countries' stock markets and show that the synchronicity of different stock markets is associated to their betweenness centrality measure. Synchronicity may also be negatively correlated to innovation. Katila and Chen [11] show that companies that introduce more innovative products are less likely to sync with other companies in their search for innovation.

In online social media, the productivity metric for a user can be the degree that the user generate and/or disseminate valuable information. One aspect of the "value" is how much the user can influence others. There has been extensive research on social influence in social media [3, 8]. Another aspect is how much is the user's information can be trusted (*e.g.*, not spams or disinformation) [14]. Researchers have proposed approaches to identify spammers based on their behavior [4, 15].

In contrast to previous studies, which mostly concentrate on social networks' static attributes or non-structural dynamic attributes, our study focuses on the inter-dependent dynamic behavior of heterogeneous objects in relation to their productivity measures. In particular, we propose computational models for predicting the productivity of teams, individuals and information in organizational settings.

2.2 Composite Network Analysis Recently, researchers have also started to address the challenges of interconnected real-world, multiple-typed objects which form heterogeneous networks. For such networks, Sun *et. al.* [20] propose a user-guided clustering approach, and Ji *et. al.* [9] design a combined ranking and classification method. Heterogeneous networks are often dynamic where links can be constantly added or deleted. Aggarwal *et. al.* study the problem of predicting links' dynamic addition and deletion [2].

The multiple-faceted relationship among the heterogeneous objects can also assist to tackle practical problems. The three different relation models among member, group and product help to improve spam detection in online product reviews [15]. Similarly, the heterogeneous relation among textual and multimedia data can improve the understanding the semantics [16]. In contrast, our study focuses on revealing the relationship between heterogeneous objects' dynamic behavior and their value to improve people's productivity, and understanding the structure within their synchronicity.

3 Datasets and Scenarios

In this section, we describe data characteristics for the two scenarios where our approach is evaluated.

3.1 Enterprise Dataset In the first scenario, our goal is to find entities in enterprises that may have high productivity (*e.g.*, teams, individuals and information) from observations in enterprise social networks. Towards this goal, we collect data about the electronic communication social networks inside a global information technology firm with more than 400,000 employees. The data contain people's electronic communication and employee job-related information.

The people's electronic communication data include email and instant messaging activities. We collected detailed electronic communication records of 8952 volunteer employees in more than 70 countries. To preserve privacy, the original textual content of an email or text message is not saved. Instead, the content is represented as a vector containing the terms appeared in the text as well as their counts after stop-words removal and stemming. In addition, to construct a view of the network that reflects the real communications, we eliminate spam and mass email announcements and are left with 20 million emails and text messaging chats. We extract 200 latent topics from the textual content, using Latent Dirichlet Allocation (LDA) [5]. From the communication data, we derive the social networks of employees within the firm. The social distance between employee i and j is defined by considering the degree of separation in their communication and the amount of communication [12, 23, 21]. Specifically, it is

$$dist(i,j) = \sum_{k=1}^{K-1} \frac{1}{strength(v_k, v_{k+1})} \tag{3.1}$$

where $v_1, ..., v_k$ are the nodes on the shorted path from user i to user j, and $strength(v_k, v_{k+1})$ measures the strength of communications between v_k and v_{k+1} and is normalized to vary between 0 and 1. The measure is defined as:

$$strength(i,j) = \frac{log(X_{ij})}{\max_j log(X_{ij})} \tag{3.2}$$

where X_{ij} is the total communications between user i and j (If $X_{ij} < 10$, we set $X_{ij} = 10$).

We also collected detailed financial performance data of more than $10,000$ consultants. These consultants generate revenue by logging "billable hours". Previous study has found that a consultant ability to generate revenue is the most appropriate productivity measure [23]. To normalize across consultants in different geographical locations and career levels with different cost base, we choose "utilization rate" as the performance metric. It has been an important metric to evaluate consultants for promotion and pay raises. The "utilization rate" is defined as the number of billable hours divided by the total number of work hours in a particular time period. In our study, the time period is from June, 2007 to July, 2008. Combining the financial data with communication data yields a total of 1029 consultants whom we have both network and financial performance data. We also record the projects which these consultants worked on. Based on the projects, the consultants can be grouped into 265 teams. To protect the privacy of the participants, their identities are replaced with hash identifiers. To match the timing of the content and the performance data, we apply the same time window of the performance data for the content, and are left with 2.1 million emails and text messaging chats. In addition, we looked up the positions of the consultants in the enterprise organizational hierarchy, as well as the time zones where their offices are located.

3.2 Social Media Dataset In addition to the enterprise scenario, our approach can be applied to finding valuable information sources on social media. Compared to traditional media, social media such as Twitter is becoming a faster channel to obtain high value information about emerging events. However, there is much more noise such as spam in Twitter. Therefore, it is necessary to assess different users' quality as information sources.

For a preliminary validation of our approach in social media context, we use a subset of annotated Twitter dataset from [4]. It contains $70K$ tweets from 344 Twitter user accounts in 2010. The accounts are manually verified and categorized into four groups: celebrities, information disseminators such as traditional media's Twitter accounts, normal users and spammers. Our goal is to automatically find the users providing high value information (*e.g.*, celebrities and information disseminators) based on users' dynamic behavior.

4 Synchronicity Networks

Previous studies in biology and sociology have shown that synchronous behavior enhances individual and group performance. For business settings, Saavedra *et. al.* [18] recently found that the higher degree the traders' synchronous activity patterns are, the less likely they may lose money. The benefit of synchronicity can come from two aspects. First, individuals facing uncertainties or risks tend to follow the wisdom of crowds. Thus, higher synchronicity of a person may mean that s/he is more effective in leveraging the collective wisdom. Second, high synchronicity arises from efficient collaboration without the aid of centralized leadership. Therefore, synchronicity is a useful feature to characterize collaboration, especially when actual collaboration activities (*e.g.*, face-to-face casual meetings) are not completely recorded.

In [18], a synchronicity metric for trader i at time period t is defined based on the number of traders with simultaneous trading as him/her. In our work, we consider the synchronicity in the information exchange activities of information workers. The information in their communication is categorized as a set of latent topics. For person i, his/her topics during time period t (*e.g.*, a given day) can be represented by a vector

$$\vec{o}_{it} =< o_{i1t}, o_{i2t}, ..., o_{ikt}, ..., o_{iMt} > \tag{4.3}$$

where o_{ikt} is the normalized strength of the k-th topic and M is the number of topics. Here o_{ikt} is calculated by dividing the topic coefficients in Latent Dirichlet Allocation [5] by the maximal topic coefficients of person i over the whole time period. In a similar spirit to [18], we define the synchronicity metric for person i on topic k at time period t as:

$$s_{ikt} = T_{ikt} - T^*_{ikt} \tag{4.4}$$

where T_{ikt} is calculated as:

$$T_{ikt} = o_{ikt} \cdot \Sigma_{j=1}^{N} o_{jkt} \tag{4.5}$$

and T^*_{ikt} is the average of T_{ikt} across an randomized ensemble within which the activities of person i were randomly shuffled.

Then, a vector of synchronicity metric can be obtained by averaging over all time periods:

$$\vec{s}_i =< s_{i1}, s_{i2}, ..., s_{ij}, ..., s_{iM} > \tag{4.6}$$

where $s_{ik} = \texttt{avg}_t(s_{ikt})$. In our experiments, however, we do not observe strong relationship between synchronicity metric $\vec{s}_i$ and the "utilization rate" of consultant i (see details in Section 6). The lack of strong relationship can be attributed to several reasons. First, unlike financial traders whose single most important activity is trading, consultants in enterprises are involved in more diverse activities such as business discussions, presentations, technical development and *etc*. Thus, synchronicity in some activities may be less beneficial or even harmful to performance than others. Second, the consultants are in much larger social networks, where there can be complex informal network structure (because of communication) and formal network structure (due to management hierarchies). In this context, synchronicity with arbitrary groups of people may not necessarily improve a consultant's performance. Instead, only synchronizing with appropriate people will help. By analogy to electronic circuits, the synchronization among people (*i.e.*, nodes) need to be carefully designed to ensure the performance of the whole network.

To address these issues, we propose to model the structure of synchronicity using *synchronicity networks*. Our hypothesis is that high performers will selectively synchronize with other high performers to achieve efficient collaboration. The nodes of the synchronicity networks can be people, and the edge between two people i and j indicates that there are synchronous activities between them. The weight of the edge for topic k at time period t is defined as:

$$e_{ij(kt)} = o_{ikt} \cdot o_{jkt} \tag{4.7}$$

Note the synchronicity network is dynamic where the edge weight changes over time. To investigate people's aggregated performance in synchronicity networks, we can use the aggregated edge weights as $e_{ij(k)} = \Sigma_t e_{ij(kt)}$. We may also aggregate edge weight over all topics as

$$e_{ij} = \Sigma_k e_{ij(k)} \tag{4.8}$$

The constructed synchronicity networks can then be used to improve finding valuables. Based on prior research's results [4, 22], we first use Support Vector Regression (SVR) [10] to build value predictors based on static features. For the enterprise data, we use the features identified in [10], including network constraint, network size and strong links to managers. For the Twitter data, we utilize the features on users' interaction [4], including the frequency of tweets and "@-mentions". Then, we use the synchronicity networks to iteratively improve the initial value prediction. In particular, we adopt the formulation of eigenvector centrality [19, 6]:

$$r(i) = \frac{1}{\lambda}\Sigma_j[e_{ij} \cdot r(j)] \tag{4.9}$$

where $r(i)$ is the value of node i and e_{ij} is the edge weight between i and j. Equation 4.9 can also be expressed as an update equation in matrix format:

$$R \leftarrow E \cdot R \tag{4.10}$$

where E is the adjacency matrix containing the edge weights, and we omit the constant λ. In addition to the people's network, we also need to find valuable teams and topics in the multi-level composite networks illustrated in Figure 1. The detailed algorithm will be discussed in Section 5.

5 Ranking Heterogeneous Objects in Multi-Level Composite Synchronicity Networks

As Figure 1 illustrates, heterogeneous objects such as teams, individuals and topics are inter-related. When a high performing team emerges, enterprises are usually interested in understanding how the team grows to be valuable. For example, how the team is composed and what information the team members tap info. Moreover, facing continuously evolving information, it is desirable for enterprises to identify what topics may benefit which individuals and teams.

5.1 Heterogeneous Multi-Level Synchronicity Networks We propose to construct a multi-level composite synchronicity network towards this goal. Let G denote the multi-level network, and G_n denote the network at the n-th level. For the network in Figure 1, we label the levels from 0 to 2 in a bottom-up way, with level 0 being the topic level network. The network can be extended to have more levels. For example, on higher levels, there can be corporate-level or societal-level networks that describe the synchronicity among different corporations or societies. On lower levels, there can be term-level network describing the relationship among terms in a topic.

Within each level of the network, the edges indicate the synchronicity relationships. Let $E^{(n)}$ denote the adjacency matrix that contains the edge weights at the n-th level. At the team-level network ($n = 2$), the edge between two teams I and J is the aggregated synchronicity between all pairs of individuals in the two teams. The weight is computed as:

$$e^{(2)}_{IJ} = \Sigma_{i \in I, j \in J}(e^{(1)}_{ij}) \tag{5.11}$$

where $e^{(1)}_{ij}$ is an element of $E^{(1)}$, and is the e_{ij} defined in Equation 4.8. Similarly, the edge between two topics in topic-level network indicates the synchronicity between the topics' dynamics. For the k-th topic at time period t, vector $\vec{p}_{kt} =< o_{1kt}, o_{1kt}, ..., o_{ikt}, ..., o_{Nkt} >$ shows its strength across different people, where o_{ikt} is defined in Equation 4.3. For $E^{(0)}$, we compute the synchronicity between the the k-th topic and m-th topic as:

$$e^{(0)}_{km} = \Sigma_t[\cos(\vec{p}_{kt}, \vec{p}_{mt})] \tag{5.12}$$

where $\cos(\vec{p}_{kt}, \vec{p}_{mt})$ is the cosine similarity between $\vec{p}_{kt}$ and $\vec{p}_{mt}$.

For cross-level edges, the edges between the team and individual levels indicate "belong to" relationship, and the edges between the individual and topics levels indicate "talk about" relationship. Let $A^{(n)}$ denote the adjacency matrix containing the weights of the cross-level edges between level n and level $(n-1)$. For $A^{(2)}$, we define the weight of the edge between the team I and individual j as:

$$a^{(2)}_{Ij} = \begin{cases} 1 : & \texttt{if } j \in I \\ 0 : & \texttt{otherwise} \end{cases}$$

Note that the teams in our scenario are project teams. Therefore, an individual may be in different project teams during different time period. For the edge between the i-th individual consultant and k-th topic, the weight in $A^{(1)}$ is defined as:

$$a^{(1)}_{ik} = \Sigma_t(o_{ikt}) \tag{5.13}$$

where o_{ikt} is the normalized strength of the k-th topic in the i-th consultant communication during time period t, as defined Equation 4.3. The cross-level adjacency matrices can be utilized to improve the value prediction, in a similar spirit to Equation 4.10.

$$R^{(n)} \leftarrow A^{(n)} \cdot R^{(n-1)} \tag{5.14}$$

$$R^{(n)} \leftarrow (A^{(n)})^T \cdot R^{(n+1)} \tag{5.15}$$

5.2 Ranking Heterogeneous Objects Given such a heterogeneous multi-level synchronicity network, a ranking algorithm is desired to simultaneously find the valuable teams, individuals and topics. In addition to simultaneously ranking heterogeneous objects, it is important to be able to update their ranking if dynamic changes occur in the network. For example, when a certain topic suddenly gains momentum in enterprises, it is helpful to understand how certain people and teams' performance will change. Likewise, it is desirable to predict what topics will be more valuable when teams are re-organized. To achieve this goal, we propose a novel algorithm for **H**eterogeneous **M**ulti-level networks **R**anking, denoted as HMR algorithm, in a similar spirit to a ranking method for networks with three types of relationship [15]. Given a heterogeneous multi-level synchronicity network as input, the HMR algorithm outputs $R^{(n)}$, which denotes the value of the nodes at level n. Let $R_x^{(n)}$ denote the value of $R^{(n)}$ at the x iteration of HMR algorithm. First, $R^{(n)}$ is initialized by existing value predictors. For the individual-level network, $R_0^{(1)}$ can be initialized using the SVR predictor of consultants' performance mentioned in Section 4. For other levels, $R_0^{(n)}$ can be initialized with 1. Then, the initial predictions of $R^{(n)}$ are iteratively improved using the multi-level network. In each iteration, the value of nodes $R^{(n)}$ is updated three times: (1) from same level network neighbors $E^{(n)}$ (Equation 4.10); (2) from $n-1$ level neighbors $A^{(n-1)}$ (Equation 5.14); and (3) from $n+1$ level neighbors $A^{(n+1)}$ (Equation 5.15). The algorithm is illustrated below:

Algorithm 5.1. HMR Algorithm

```
Input: Adjacency matrices $E^{(n)}$, $A^{(n)}$,
       initial value vectors $R_0^{(n)}$
Output: Value vectors $R^{(n)}$

1. Initialize $x \leftarrow 0$;
2. Iterate:
     $x \leftarrow x+1$
     for $n \leftarrow 1$ to $N-1$ do
         $R_{x-1}^{(n)} \leftarrow A^{(n)} \cdot R_{x-1}^{(n-1)}$
     for $n \leftarrow 0$ to $N-2$ do
         $R_{x-1}^{(n)} \leftarrow (A^{(n)})^T \cdot R_{x-1}^{(n+1)}$
     for $n \leftarrow 0$ to $N-1$ do
         $R_x^{(n)} \leftarrow E^{(n)} \cdot R_{x-1}^{(n)}$
     $R_x^{(n)} \leftarrow \frac{R_x^{(n)}}{\|R_t^{(x)}\|_1}$
   until $\| R_x^{(n)} - R_{x-1}^{(n)} \|_1 < \delta$
3. Return $R_x^{(n)}$;
```

Here δ is a constant chosen by users to control the number of iterations. When there are dynamic changes of nodes, HMR algorithm can propagate the changes to inter-related nodes to get updated value vectors.

The convergence of the HMR algorithm can be proved by showing that HMR is an instance of an eigenvalue problem. From line 2 of Algorithm 5.1, we have

$$R_{x-1}^{(n)} = (A^{(n)})^T \cdot A^{(n)} \cdot R_{x-1}^{(n)} \tag{5.16}$$

$$R_x^{(n)} = E^{(n)} \cdot R_{x-1}^{(n)} \tag{5.17}$$

Combining the two equations, we have

$$R_x^{(n)} = [E^{(n)} \cdot (A^{(n)})^T \cdot A^{(n)}] \cdot R_{x-1}^{(n)} \tag{5.18}$$

Let $B = E^{(n)} \cdot (A^{(n)})^T \cdot A^{(n)}$. It can be seen that this is an instance of power iteration for the eigenvalue problem of computing $R^{(n)}$ as the eigenvector of matrix B corresponding to the dominant eigenvalue. Therefore, the HMR algorithm seeks to align $R^{(n)}$ towards the dominant eigenvector of B. It can be shown to converge following a similar proof in [15].

6 Experiment

In this section, we evaluate the effectiveness of the proposed heterogeneous multi-level synchronicity networks and the HMR algorithm for finding valuable objects. We first validate our approach using the multi-faceted employee productivity data collected at a large international enterprise. Next, we present the preliminary evaluation of our approach in social media domain using Twitter data.

6.1 Enterprise Productivity Data

As described in section 3.1, the enterprise productivity data contain 265 project teams, 1032 consultants and 200 extracted latent topics. To evaluate our approach, we use the following ground truth of their values. For individual consultants, the ground truth is individual "utilization rate". For teams, the ground truth is the group "utilization rate" (*i.e.*, the mean of individual "utilization rate" of team members). For topics, it is the topic value indicator described in [22], which is the coefficients of topics in the linear regression of productivity on topics.

6.1.1 Empirical Analysis

To ensure that the proposed synchronicity and synchronicity network features are good indicators of productivity, we first analyze them by validating their relationship with consultants' productivity measures.

We first use the difference between two consultants' time zones as a proxy to study their synchronicity. The rationale is that it is more difficult for people to synchronize in their activities if there are large differences in their time zones. Conversely, previous work [7, 23] has shown that people can improve their productivity by building diverse social networks. Collaborating with people in different time zones will increase diversity of a consultant's social network. Therefore, the two fac-

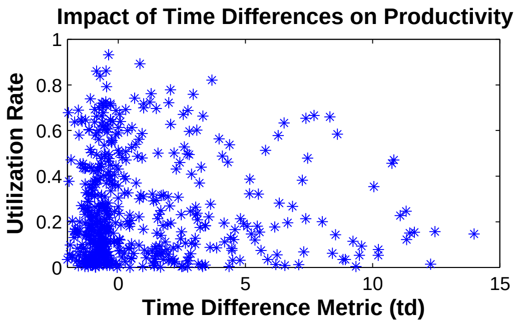


Figure 2: The impact of time zone differences in employees' social networks on their productivity.

tors, synchronicity and diversity, may be inherently conflicting with each other. We would like to understand whether synchronicity still significantly impact consultants' performance under such conditions. We define a time difference metric in a person's social network as:

(6.19)
$$td(i) = \log(\Sigma_{j \in D_1(i)}[|\mathtt{tz}(i) - \mathtt{tz}(j)| \cdot \exp(-dist(i,j))])$$

where $D_1(i)$ is the 1-degree neighbor set of consultant i, $\mathtt{tz}(i)$ is the time zone where consultant i's office is located and $dist(i,j)$ is the social distance between i and j defined in Equation 3.1.

The relationship between consultants' utilization rate and their time difference metric is illustrated in Figure 2. We can observe that most high performers have low time difference metric values. In contrast, only a few consultants have high time difference metric values, and most of them have low utilization rate (*e.g.*, < 0.5). The linear regression of utilization rate on $td(i)$ gives a correlation coefficient of -0.10 ($p < 0.1$). Thus, the result indicates that synchronicity does impact productivity.

In addition, we perform linear regression to study the relationship between synchronicity metric $\vec{s}_i$ and "utilization rate". To calculate $\vec{s}_i$, we need to select an appropriate unit for synchronicity time period t (Equation 4.3), within which the activities are considered synchronous. We experimented with time unit values equal to one day, one hour, and one minute. Only the regression with a time unit of one hour yields statistically significant results. We believe that this is because the delay of corporate emails in coordinated work activities is usually between one minute and one hour. Note that the appropriate time unit value for t can be application dependent. For example, the time unit value for stock traders can be one second [18]. For the time unit of one hour, the R-square statistics equals to 0.31 and $p < 0.03$, suggesting that synchronicity indeed have statistically significant effects on productivity. But the small R-square statistics shows the relationship is not strong enough to reliably predict productivity from synchronicity alone.

	team	consultant	topic
sync-HMR	65.1%	70.0%	71.4%
static-HMR	57.3%	64.7%	67.2%
sync-SVR	55.4%	57.8%	52.5%
static-SVR	54.2%	54.6%	53.9%

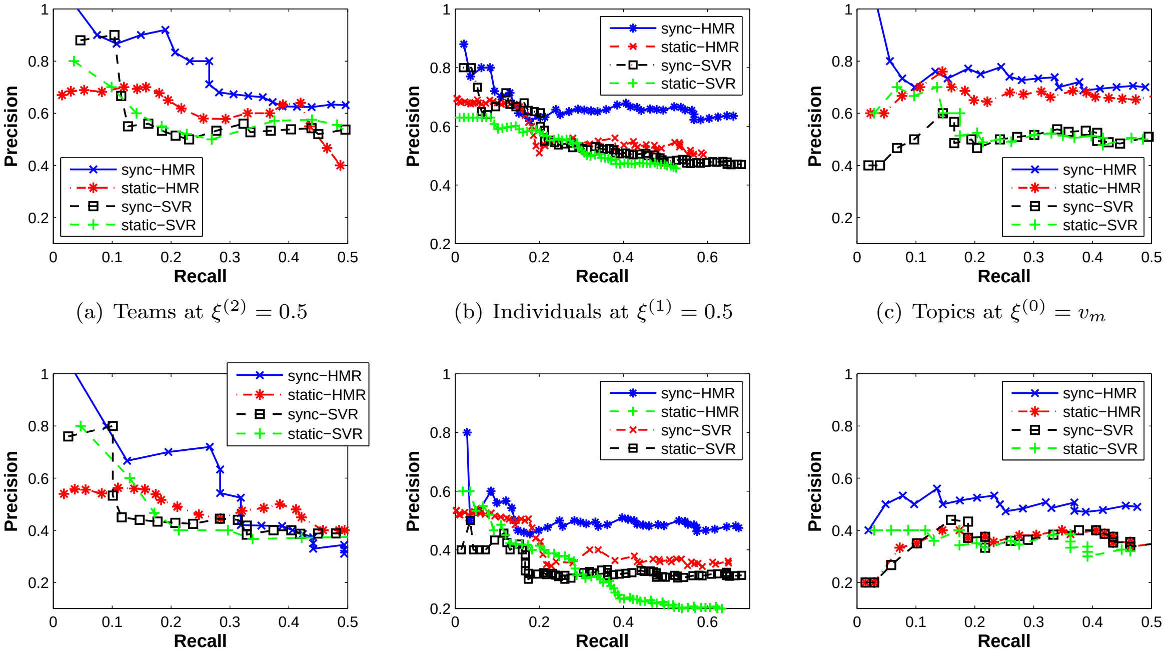
Table 1: The percentage of correct pairwise comparisons.

6.1.2 Ranking Results

To evaluate the effectiveness of HMR algorithm in ranking heterogeneous objects in multi-level synchronicity networks (denoted as *sync-HMR*), we use the following three baseline algorithms. The first baseline is called *static-HMR*, where we apply HMR algorithm on multi-level networks constructed based on people's communication. The edge weights between teams and individuals are defined similarly as Equation 3.2, and the edge weight between two topics is decided by the cosine similarity. The second baseline is *sync-SVR*, where the object value is predicted by SVR using synchronicity feature $\vec{s}_i$ defined in Equation 4.6 and other features from prior research including network constraint, network size and strong links to managers (see section 4). The last baseline is *static-SVR* which is similar to sync-SVR except that the synchronicity feature is not used. All results of the four methods are obtained using five-fold cross validation.

To evaluate the rankings of the four methods, we first calculate the percentage that each method gives correct results in all pairwise comparisons. Such pairwise comparisons are useful in practice because it is often required to compare two teams (or individual, topics) to see which one is more valuable to improve enterprise productivity. The results on the percentage of correct pairwise comparisons for the four methods are listed in Table 1. We can observe that *sync-HMR* significantly outperforms the three baseline methods.

Besides pairwise comparisons, users in many applications may want to investigate a certain number of objects that are highly likely to be valuable. Thus, we use precision and recall at top K to evaluate the rankings. In this case, we use a threshold $\xi^{(n)}$ to classify high value and low value objects. The true positive high value objects are those whose ground truth value $\geq \xi^{(n)}$. For individual consultants, the threshold $\xi^{(1)}$ is on individual "utilization rate". For teams, the threshold $\xi^{(2)}$ is on the group "utilization rate". For topics, the threshold $\xi^{(0)}$ is the topic value indicator. In our experiment, we use two sets of threshold values: (T1)



(a) Teams at $\xi^{(2)} = 0.5$ (b) Individuals at $\xi^{(1)} = 0.5$ (c) Topics at $\xi^{(0)} = v_m$

(d) Teams at $\xi^{(2)} = 0.75$ (e) Individuals at $\xi^{(1)} = 0.75$ (f) Topics at $\xi^{(0)} = 1.5v_m$

Figure 3: The precision-recall curves for two sets of threshold, where v_m is the the median of topic value indicator.

$\xi^{(1)} = \xi^{(2)} = 0.5$, $\xi^{(0)}$ = the median of topic value indicator; and (T2) $\xi^{(1)} = \xi^{(2)} = 0.75$, $\xi^{(0)} = 1.5$ times the median of topic value indicator. For given threshold $\xi^{(n)}$ values, we compute the precision and recall at various K values ($K = 5, 10, 15, ..., 100$). The corresponding precision recall values are plotted in Figure 3. We can see that our approach *sync-HMR* consistently outperforms all baseline methods. It can also be observed that *sync-SVR* does not always outperform *static-SVR*, suggesting the importance of structural features for synchronicity.

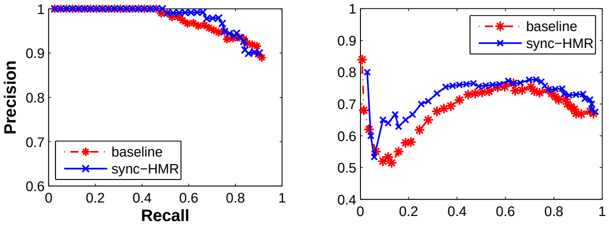
6.2 Twitter data In addition to the enterprise dataset, we evaluate our approach in a Twitter dataset from [4] (see Section 3.2) to understand whether synchronicity networks can improve finding valuables in social media domain.

In this preliminary study, we choose to first focus on the synchronicity networks of individual twitter accounts, since there is no explicit team information available and it is challenging to apply LDA type of topic modeling on tweets. Thus, we ignore the differences in topics when computing synchronicity. Instead, we assume a general topic for all tweets in Equation 4.3 to Equation 4.8.

To evaluate our approach, we use the manually verified group label of each user (*i.e.*, celebrities, information disseminators, normal users, or spammers) as the ground truth. Furthermore, we assign the value rank of 3 to 0 to the four groups, where celebrities have rank 3 and are the most valuable ones. We also choose the algorithm in [4] as the baseline. The baseline algorithm uses PageRank to compute user reputation on the network extracted from users' "@-mentions" interactions weighted by their tweeting behavior metrics. First, in the pairwise comparisons, snyc-HMR is correct 90.2% of the time while the baseline is 86.9%. In addition, we calculate precision and recall at top K ($K = 5, 10, ..., 100$) for threshold $\xi = 1$ and $\xi = 2$. $\xi = 1$ corresponds to the application to filter out spammers. $\xi = 2$ means that it detects celebrities and information disseminators. The results are illustrated in Figure 4. It can be observed that our approach also outperforms the baseline method, although the baseline is already very good at filtering spammers.

7 Conclusion

In this paper, we propose a novel approach for finding heterogeneous entities that are valuable to improve people's productivity in organizational networks. First, our approach uses multi-level synchronicity networks to characterize the structures of the heterogeneous entities' *dynamic* interactions. The synchronicity networks pro-



(a) (b)

Figure 4: (a) The precision-recall curves for $\xi = 1$. (b) The precision-recall curves for $\xi = 2$..

vide insights on the entities' ability to leverage collective wisdom and thus can be used to predict productivity. We then propose a novel algorithm for **H**eterogeneous **M**ulti-level networks **R**anking, denoted as HMR algorithm. The HMR algorithm can effectively find interrelated high value heterogeneous entities, such as top individual performers, the most useful information they access and the best teams they have been involved. Experimental results on enterprise performance data set show that our approach significantly outperforms exiting methods that do not use structural features of dynamic interactions. In addition, our preliminary study demonstrates that our approach can improve categorization of Twitter users over previous methods.

References

[1] A. Adarov, R. Kali, and J. Reyes. Stock market synchronicity and the global trade network: A random-walk approach. Technical Report FREIT115, University of Arkansas, 2009.

[2] C. C. Aggarwal, Y. Xie, and P. S. Yu. On dynamic link inference in heterogeneous networks. In *SDM*, pages 415–426, 2012.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, 2011.

[4] V. A. Balasubramaniyan, A. Maheswaran, V. Mahalingam, M. Ahamad, and H. Venkateswaran. A crow or a blackbird?: Using true social network and tweeting behavior to detect malicious entities in twitter. Technical Report GT-CS-10-14, Georgia Institute of Technology, 2010.

[5] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of Inter. WWW Conf.*, 1998.

[7] R. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1992.

[8] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.

[9] M. Ji, J. Han, and M. Danilevsky. Ranking-based classification of heterogeneous information networks. In *KDD*, pages 1298–1306, 2011.

[10] T. Joachims. *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. B. Schlkopf and C. Burges and A. Smola. (ed.). The MIT Press, 1999.

[11] R. Katila and E. L. Chen. Effects of search timing on innovation: The value of not being in sync with rivals. *Administrative Science Quarterly*, pages 593–625, December 2008.

[12] C. Lin, K. Ehrlich, V. Griffiths-Fisher, and C. Desforges. Smallblue: People mining for expertise search,. *IEEE Multimedia Magazine*, 15(1):78–84, 2008.

[13] C.-Y. Lin, L. Wu, Z. Wen, H. Tong, V. Griffiths-Fisher, L. Shi, and D. Lubensky. Social network analysis in enterprise. *Proceedings of the IEEE*, 100:2759 – 2776, 2012.

[14] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 71–79, 2010.

[15] A. Mukherjee, B. Liu, and N. S. Glance. Spotting fake reviewer groups in consumer reviews. In *WWW*, pages 191–200, 2012.

[16] G.-J. Qi, C. C. Aggarwal, and T. S. Huang. Transfer learning of distance metrics by cross-domain metric sampling across heterogeneous spaces. In *SDM*, pages 528–539, 2012.

[17] S. Rodan and D. Galunic. More than network structure: How knowledge heterogeneity influences managerial performance and innovativeness. *Strategic Management Journal*, 25:541–556, 2004.

[18] S. Saavedra, K. Hagerty, and B. Uzzi. Synchronicity, instant messaging, and performance among financial traders. *Proceeding of the National Academy of Sciences (PNAS)*, 108:5296–5301, March 2011.

[19] G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31:581–603, 1966.

[20] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In *KDD*, pages 1348–1356, 2012.

[21] Z. Wen and C.-Y. Lin. On the quality of inferring interests from social neighbors. In *KDD*, pages 373–382, 2010.

[22] Z. Wen and C.-Y. Lin. Toward finding valuable topics. In *SDM*, pages 720–731, 2010.

[23] L. Wu, C. Lin, S. Aral, and E. Brynjolfsson. Value of social network – a large-scale analysis on network structure impact to financial revenue of information technology consultants. In *The Winter Conference on Business Intelligence*, 2009.