Information Diffusion at Workplace

Jiawei Zhang*, Philip S. Yu*,¶, Yuanhua Lv†, Qianyi Zhan‡
*University of Illinois at Chicago, Chicago, IL, USA
¶Institute for Data Science, Tsinghua University, Beijing, China
†Microsoft Research, Redmond, WA, USA
‡Nanjing University, Nanjing 210023, China
jzhan9@uic.edu, psyu@cs.uic.edu, yuanhual@microsoft.com, zhanqianyi@gmail.com

ABSTRACT

People nowadays need to spend a large amount of time on their work everyday and workplace has become an important social occasion for effective communication and information exchange among employees. Besides traditional offline contacts (e.g., face-to-face meetings and telephone calls), to facilitate the communication and cooperation among employees, a new type of online social networks has been launched inside the firewalls of many companies, which are named as the "enterprise social networks" (ESNs). In this paper, we want to study the information diffusion among employees at workplace via both online ESNs and offline contacts. This is formally defined as the IDE (Information Diffusion in Enterprise) problem. Several challenges need to be addressed in solving the IDE problem: (1) diffusion channel extraction from online ESN and offline contacts; (2) effective aggregation of the information delivered via different diffusion channels; and (3) communication channel weighting and selection. A novel information diffusion model, **Muse** (Multi-source Multi-channel Multi-topic diffUsion SElection), is introduced in this paper to resolve these challenges. Extensive experiments conducted on real-world ESN and organizational chart dataset demonstrate the outstanding performance of **Muse** in addressing the IDE problem.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining

Keywords

Diffusion Channel Selection, Enterprise Social Networks, Data Mining

1. INTRODUCTION

On average, people nowadays need to spend more than 30% of their time at work everyday. According to the statistical data in [7], the total amount of time people spent at workplace in their life is tremendously large. For instance, a young man who is 20 years old now will spend 19.1% of his future time working [7]. Therefore, workplace is actually an easily neglected yet important social occasion for effective communication and information exchange among people in our social life.

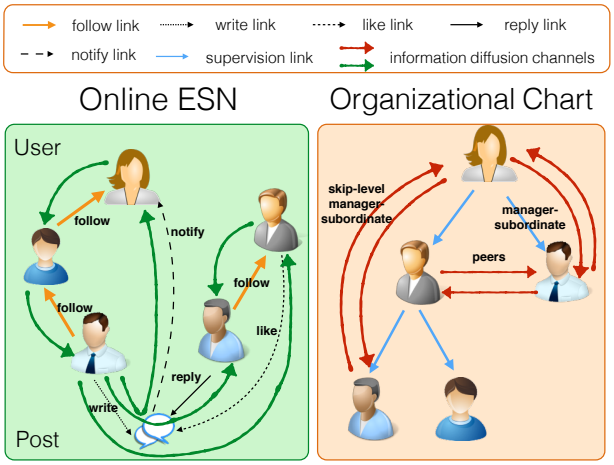


Figure 1: An example of information diffusion at workplace.

Besides the traditional offline contacts, like face-to-face communication, telephone calls and messaging, to facilitate the cooperation and communications among employees, a new type of online social networks named Enterprise Social Networks (ESNs) has been launched inside the firewalls of many companies [22, 21]. A representative example is Yammer, which is used by over 500,000 leading businesses around the world, including 85% of the Fortune 500[1]. Yammer provides various online communication services for employees at workplace, which include instant online messaging, write/reply/like posts, file upload/download/share, etc. In summary, the communication means existing among employees at workplaces are so diverse, which can generally be divided into two categories [18]: (1) offline communication means, and (2) online virtual communication means.

Problem: In this paper, we will study how information diffuses via both online and offline communication means among employees at workplace, which is formally defined as the "Information Diffusion in Enterprise" (IDE) problem.

[1]https://about.yammer.com/why-yammer/

CIKM'16 , October 24-November 28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983848

Table 1: Summary of related problems.

Property	Information Diffusion at Workplace	Social Network Information Diffusion [1]	Organization Hierarchy Information Diffusion [3]	Multi-Network Influence Max. [13]
#diffusion sources	multiple (ESN + Chart + Hybrid)	single	single	multiple
network types	heterogeneous + tree	heterogeneous	tree	homogeneous
#channels per source	multiple	single	single	single
#topics	multiple	single	single	single

To help illustrate the IDE problem more clearly, we also give an example in Figure 1. The left plot of Figure 1 is about an online ESN, employees in which can perform various social activities. For instances, employees can follow each other, can write/reply/like posts online, and posts written by them can also @certain employees to send notifications, which create various online information diffusion channels (i.e., the green lines) among employees. Meanwhile, the relative management relationships among the employees in the company can be represented with the organizational chart (i.e., the right plot), which is a tree-structure diagram connecting employees via supervision links (from managers to subordinates). Colleagues who are physically close in the organizational chart (e.g., peers, manager-subordinates) may have more chance to meet in the offline workplace. For example, subordinates need to report to their managers regularly, peers may co-operate to finish projects together, which can form various offline information diffusion channels (i.e., the red lines) among employees at workplace.

The IDE problem is an important problem. By studying the IDE problem, we can gain a better understanding about how information propagates in workplace, which will lead to lots of benefits for both employees and companies (e.g., *more efficacious communications among employees* and *broader the information diffusion in companies*). By addressing the IDE problem, employees can choose the most effective and efficient channels to communicate with colleagues in their future work, which can improve their work efficiency greatly. Via a combination of several communication channels, companies can convey important messages, e.g., recent organizational changes, new employees and new products, to all the employees in the company quickly.

Besides its importance, the IDE problem is a novel problem and totally different from existing works on information diffusion and influence maximization: (1) "*information diffusion in social networks*" [1, 6], which study the diffusion of information among users in social networks only; (2) "*technology adoption via organization hierarchy*" [3], which focuses on the information diffusion in/between management levels in companies; and (3) "*influence maximization in multiple social networks*" [13], which explores the influence maximization problem in multiple homogeneous social networks by simply merging them into one single network. Different from all these related works, in this paper, we aim at studying how information propagates via various communication channels in both online ESNs and offline organizational chart at workplace. A more detailed comparison of IDE with these related problems is available in Table 1.

Despite its importance and novelty, the IDE problem is very hard to address due to the following challenges:

- *Diffusion Channel Extraction and Inference*: Online ESNs provide employees with various communication services, which can create different online information diffusion channels among employees. Meanwhile, employees' offline contacts belong to their personal private information, which are confidential to both companies and the public. As a result, extracting the diffusion channels from the heterogeneous online ESNs and inferring potential offline diffusion channels among employees in companies are both very challenging.
- *Diffusion Channel Aggregation*: Employees in workplaces can receive information from their neighbors via diverse channels. New information diffusion models which can aggregate all these channels is required to depict the information diffusion process at workplace.
- *Diffusion Channel Weighting and Selection*: Different diffusion channels play different roles in delivering information among employees at workplace, some of which are helpful but some can be useless. Weighting and selecting useful channels will help improve the communication quality among employees at workplace significantly.

To address all the above challenges, a novel information diffusion model **Muse** (Multi-source Multi-channel Multi-topic diffUsion SElection) is proposed in this paper. **Muse** extracts and infers sets of online, offline and hybrid (of online and offline) diffusion channels among employees across online ESN and offline organizational structure. Information propagated via different channels can be aggregated effectively in **Muse**. Different diffusion channels will be weighted according to their importance learned from the social activity log data with optimization techniques and top-K effective diffusion channels will be selected in **Muse** finally.

The rest of this paper is organized as follows. In Section 2, we introduce some concept definitions. In Section 3, the **Muse** diffusion model will be introduced in details, which will be evaluated in Section 4. Finally, we will talk about the related works in Section 5 and conclude this paper in Section 6.

2. TERMINOLOGY DEFINITION

In this section, we will introduce the definition of several important concepts used in this paper, which include the online enterprise social networks and the organizational chart.

Definition 1 (Enterprise Social Networks (ESNs)): Online *enterprise social networks* are a new type of online social networks used in enterprises to facilitate employees' communications and daily work, which can be represented as *heterogeneous information networks* $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \bigcup_i \mathcal{V}_i$ is the set of different kinds of nodes and $\mathcal{E} = \bigcup_j \mathcal{E}_j$ is the union of complex links in the network.

In this paper, we will use Yammer as an example of online ESNs. As introduced in Section 1, users in Yammer can have various social activities, e.g., follow other users, join groups and write/reply/like posts and posts can @users to send notifications. As a result, Yammer can be represented as $G = (\mathcal{V}, \mathcal{E})$, where node set $\mathcal{V} = \mathcal{U} \cup \mathcal{O} \cup \mathcal{P}$ and $\mathcal{U}$, $\mathcal{O}$ and

$\mathcal{P}$ are the sets of users, groups and posts respectively; link set $\mathcal{E} = \mathcal{E}_s \cup \mathcal{E}_j \cup \mathcal{E}_w \cup \mathcal{E}_r \cup \mathcal{E}_l$ denoting the union of social, group membership, write, reply and like links in Yammer respectively. In this paper, we regard different group participation as the target activity, information about which can diffuse among employees at the workplace. Groups in ESNs are usually of different themes (e.g., new products, state-of-art techniques, daily-life entertainments), which are treated as different information topics in this paper.

Definition 2 (Organizational Chart): *Organizational chart* is a diagram outlining the structure of an organization as well as the relative ranks of employees' positions and jobs, which can be represented as a rooted tree $C = (\mathcal{N}, \mathcal{L}, root)$, where $\mathcal{N}$ denotes the set of employees and $\mathcal{L}$ is the set of directed *supervision links* from managers to subordinates in the company, *root* usually represents the CEO by default.

Each employee in the company can create exactly one account in Yammer with valid employment ID, i.e., there is *one-to-one* correspondence between the users in Yammer and employees in the organization chart. For simplicity, in this paper, we assume the user set in online ESN to be identical to the employee set in the organizational chart (i.e., $\mathcal{U} = \mathcal{N}$) and we will use "Employee" to denote individuals in both online ESN and offline organizational chart by default.

3. PROPOSED METHOD

3.1 Preliminary

In this section, a novel information diffusion model MUSE will be proposed to depict the information propagation process of multiple topics via different diffusion channels across the online and offline worlds at workplace. We denote the set of topics diffusing in the workplace as set $\mathcal{T}$. Three different diffusion sources will be our main focus in this paper: online source, offline source and the hybrid source (across online and offline sources). The diffusion channel set of all these three sources can be represented as $\mathcal{C}^{(on)}$, $\mathcal{C}^{(off)}$ and $\mathcal{C}^{(hyb)}$ respectively, whose sizes are $\left|\mathcal{C}^{(on)}\right| = k^{(on)}$, $\left|\mathcal{C}^{(off)}\right| = k^{(off)}$, $\left|\mathcal{C}^{(hyb)}\right| = k^{(hyb)}$.

In MUSE, a set of users are activated initially, whose information will propagate in discrete steps within the network to other users. Let v be an employee at workplace who has been activated by topic $t \in \mathcal{T}$. For instance, at step τ, v will send a amount of $w^{(on),i}(v,u,t)$ information on topic t to u via the i_{th} channel in the online source (i.e., channel $c^{(on),i} \in \mathcal{C}^{(on)}$), where u is an employee following v in channel $c^{(on),i}$. The amount of information that u receives from v via all the channels in the online source at step τ can be represented as vector $\mathbf{w}^{(on)}(v,u,t) = [w^{(on),1}(v,u,t), w^{(on),2}(v,u,t), \cdots, w^{(on),k^{(on)}}(v,u,t)]$. Similarly, we can also represent the vectors of information u receives from v through channels in offline source and hybrid source as vectors $\mathbf{w}^{(off)}(v,u,t)$ and $\mathbf{w}^{(hyb)}(v,u,t)$ respectively.

Meanwhile, users in MUSE are associated thresholds to different topics, which are selected at random from the uniform distribution in range $[0,1]$. Employee u can get activated by topic t if the information received from his active neighbors via diffusion channels of all these three sources can exceed his *activation threshold* $\theta(u,t)$ to topic t,

$$f\left(\mathbf{w}^{(on)}(\cdot,u,t), \mathbf{w}^{(off)}(\cdot,u,t), \mathbf{w}^{(hyb)}(\cdot,u,t)\right) \geq \theta(u,t),$$

where aggregation function $f(\cdot)$ maps the information u receives from all the channels to u's *activation probability* in range $[0,1]$. Here, the vector $\mathbf{w}^{(on)}(\cdot,u,t) = [w^{(on),1}(\cdot,u,t), w^{(on),2}(\cdot,u,t), \cdots, w^{(on),k^{(on)}}(\cdot,u,t)]$, where $w^{(on),i}(\cdot,u,t)$ denotes the information received from all the employees u follows in channel $c^{(on),i}$, i.e.,

$$w^{(on),i}(\cdot,u,t) = \sum_{v \in \Gamma^{(on),i}_{out}(u)} w^{(on),i}(v,u,t).$$

Vectors $\mathbf{w}^{(off)}(\cdot,u,t)$ and $\mathbf{w}^{(hyb)}(\cdot,u,t)$ can be represented in a similar way. Once being activated, a user will stay active in the remaining rounds and each user can be activated at most once. Such a process will end if no new activations are possible.

Considering that individuals' *activation thresholds* $\theta(u,t)$ to topic t is is pre-determined by the uniform distribution, next we will focus on studying the information received via channels of the *online*, *offline* and *hybrid* sources and the *aggregation function* $f(\cdot)$ in details.

3.2 Online Diffusion Channel

Online ESNs provide various communication means for employees to contact each other, where individuals who have no social connections can still pass information via many other connections. Each connection among employees can form an information diffusion channel in online ESN. In this section, we propose to extract the various diffusion channels among employees based on a set of *online social meta paths* [16] extracted from the heterogeneous information in the online ESN. Before that, we first introduce the schema of enterprise social network as follows.

Based on an online ESN $G = (\mathcal{V}, \mathcal{E})$, we can represent its network schema as $S_G = (\mathcal{T}_G, \mathcal{R}_G)$, where $\mathcal{T}_G$ and $\mathcal{R}_G$ represent the sets of node types and link types in network G respectively. For example, for the Yammer network introduced in Section 2, we can define its schema as S_G, where $\mathcal{T}_G = \{Employee, Post\}$ and $\mathcal{R}_G = \{Social^{1/-1}, Write^{1/-1}, Reply^{1/-1}, Like^{1/-1}, Notify^{1/-1}\}$, where superscript -1 denotes the reverse of the corresponding link type in online ESN (group information is used as the target social activity diffusing at workplace, which will not be applied in extracting diffusion channels).

Definition 3 (Online Social Meta Path): An online meta path can be represented as a sequence $\Phi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$, where $T_i \in \mathcal{T}_G$ and $R_j \in \mathcal{R}_G$. In this paper, we are mainly concerned about meta paths starting and ending with employee nodes (i.e., $T_1 = T_k = Employee$), which are formally defined as the *online social meta paths*.

In enterprise social networks, individuals can (1) get information from employees they follow (i.e., their followees) and (2) people that their "followees" follow (i.e., 2_{nd} level followees), and obtain information from employees by (3) viewing and replying their posts, (4) viewing and liking their posts, as well as (5) getting notified by their posts (i.e., explicitly @ certain users in posts). In this paper, we propose to extract 5 different *online social meta paths* from the online ESN, whose physical meanings, representations and abbreviated notations are listed as follows:

- Followee: $Employee \xleftarrow{Social^{-1}} Employee$, whose notation is Φ_1.

- Followee-Followee: $Employee \xleftarrow{Social^{-1}} Employee \xleftarrow{Social^{-1}} Employee$, whose notation is Φ_2.
- Reply Post: $Employee \xleftarrow{Reply^{-1}} Post \xleftarrow{Write} Employee$, whose notation is Φ_3.
- Like Post: $Employee \xleftarrow{Like^{-1}} Post \xleftarrow{Write} Employee$, whose notation is Φ_4.
- Post Notification: $Employee \xleftarrow{Notify} Post \xleftarrow{Write} Employee$, whose notation is Φ_5.

The direction of the links denotes the information diffusion direction and end of the diffusion links (i.e., the first employee of the above paths) represents the target employee to receive the information. For example, Φ_1 denotes the target user receives information from his followees, while Φ_5 means that the target employee receives information from employees who have ever written posts @ the target employee.

Each of the above *online social meta path* defines a information diffusion channel among individuals in online ESN. As a result, in this paper, $\mathcal{C}^{(on)} = \{\Phi_1, \Phi_2, \Phi_3, \Phi_4, \Phi_5\}$ and $k^{(on)} = 5$ and Φ_i is identical to $c^{(on),i}$ mentioned before (denoting the i_{th} online diffusion channel). Based on each of these online social meta paths, we can extract the corresponding path instances connecting employees u and v (i.e., the concrete information diffusion traces from v to u), which can be represented as set $\mathcal{P}^{(on)}_{\Phi_i}(v \rightsquigarrow u)$, for $\forall \Phi_i \in \mathcal{C}^{(on)}$. Furthermore, let $\mathcal{P}^{(on)}_{\Phi_i}(v \rightsquigarrow \cdot)$ and $\mathcal{P}^{(on)}_{\Phi_i}(\cdot \rightsquigarrow u)$ be the sets of path instances of Φ_i going out from v and going into u respectively, with which we can define the amount of information propagating from v to u via diffusion channel $c^{(on),i} = \Phi_i$ to be

$$w^{(on),i}(v, u, t) = \frac{2\left|\mathcal{P}^{(on)}_{\Phi_i}(v \rightsquigarrow u)\right| \cdot I(v, t)}{\left|\mathcal{P}^{(on)}_{\Phi_i}(v \rightsquigarrow \cdot)\right| + \left|\mathcal{P}^{(on)}_{\Phi_i}(\cdot \rightsquigarrow u)\right|},$$

where binary function $I(v, t) = 1$ if v has been activated by topic t and 0 otherwise.

In the above definition, the proportion of information propagated from v to u via the communication channels (i.e., $w^{(on),i}(v, u, t)$) can denote how close these two users are, which depends on (1) the number of concrete diffusion path instances between them (i.e., $\mathcal{P}^{(on)}_{\Phi_i}(v \rightsquigarrow u)$); (2) the out-degree in the channel from v (i.e., $\mathcal{P}^{(on)}_{\Phi_i}(v \rightsquigarrow \cdot)$); and (3) the in-degree in the channel to u (i.e., $\mathcal{P}^{(on)}_{\Phi_i}(\cdot \rightsquigarrow u)$).

3.3 Offline Diffusion Channel

Employees' offline interactions are actually confidential to both companies and the public, which is very hard to know exactly. To infer the potential offline information diffusion channels at workplace, we propose to extract the potential information diffusion channels among individuals based on the organizational chart of the company.

Similar to online enterprise social networks, we can define the schema of the organization chart as $S_C = (\mathcal{T}_C, \mathcal{R}_C)$, where $\mathcal{T}_C = \{Employee\}$ and $\mathcal{R}_C = \{Supervision^{1/-1}\}$, based on which, we define the offline social meta path formally as follows:

Definition 4 (Offline Social Meta Path): An offline meta path can be represented as $\Omega = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$, where $T_i \in \mathcal{T}_C$ and $R_j \in \mathcal{R}_C$. In organizational chart, there exists only one type of nodes (i.e., the Employee) and offline meta paths connecting Employee nodes are named as the *offline social meta path*.

In offline workplace, the most common social interaction should happen between close colleagues, e.g., peers, manager-subordinate, and skip-level manager-subordinates, etc. The physical meaning and notations of offline social meta paths extracted in this paper are listed as follows:

- Manager: $Employee \xleftarrow{Supervision} Employee$, whose notation is Ω_1.
- Subordinate: $Employee \xleftarrow{Supervision^{-1}} Employee$, whose notation is Ω_2.
- Peer: $Employee \xleftarrow{Supervision} Employee \xleftarrow{Supervision^{-1}} Employee$, whose notation is Ω_3.
- 2nd-Level Manager: $Employee \xleftarrow{Supervision} Employee \xleftarrow{Supervision} Employee$, whose notation is Ω_4.
- 2nd-Level Subordinate: $Employee \xleftarrow{Supervision^{-1}} Employee \xleftarrow{Supervision^{-1}} Employee$, whose notation is Ω_5.

Similarly, the direction of links represents the information flow direction and the ending employees of the paths denotes the target employee, who receives information. For instance, meta path Ω_1 means that the target employee receives information from his manager, while Ω_3 denotes that the target employee receives information from his peers.

Each employees at the workplace can be influenced by both his manager as well as his subordinates (if exist) and to clarify the difference between these two different diffusion channels, we define both Ω_1 and Ω_2 (as well as Ω_4 and Ω_5). Based on the above introduced offline social meta paths, the offline diffusion channel can be represented as set $\mathcal{G}^{(off)} = \{\Omega_1, \Omega_2, \Omega_3, \Omega_4, \Omega_5\}$ and $k^{(off)} = 5$, where Ω_i denotes the i_{th} offline diffusion channel among employees.

Based on offline social meta path, e.g., Ω_i, the amount of information on topic t propagating from employee v to u can be represented as

$$w^{(off),i}(v, u, t) = \frac{2\left|\mathcal{P}^{(off)}_{\Omega_i}(v \rightsquigarrow u)\right| \cdot I(v, t)}{\left|\mathcal{P}^{(off)}_{\Omega_i}(v \rightsquigarrow \cdot)\right| + \left|\mathcal{P}^{(off)}_{\Omega_i}(\cdot \rightsquigarrow u)\right|},$$

where $\mathcal{P}^{(off)}_{\Omega_i}(v \rightsquigarrow u)$ denotes the offline social meta path instance set of Ω_i connecting v to u in the chart.

3.4 Hybrid Diffusion Channel

Besides the pure online/offline diffusion channels, information can also propagate across both online and offline worlds simultaneously. Consider, for example, two employees v and u who are not connected by any diffusion channels in online ESN or offline workplace, v can still influence u by activating u's manager via online contacts and the manager will further propagate the influence to v via offline interactions. To represent such a kind of diffusion channels, we define the concept of *hybrid social meta paths* as follows:

Definition 5 (Hybrid Social Meta Path): A hybrid meta path can be represented as $\Psi = T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_{k-1}} T_k$, where $T_i \in \mathcal{T}_G \cup \mathcal{T}_C$, $R_j \in \mathcal{R}_G \cup \mathcal{R}_C$. Meta paths which starting and ending with Employee node type are formally defined as the *hybrid social meta path*.

As proposed in [9], every pair of people in the worlds can get connected via 6 hops (i.e., six degrees of separation theory). To avoid connecting all the employees by hybrid diffusion channels, we limit its length (i.e., the number of relations in the meta path) to 3 only. The set of hybrid social meta path used in this paper, together with their physical meanings, notations are listed as follows:

- Followee-Manager: *Employee* $\xleftarrow{Social^{-1}}$ *Employee* $\xleftarrow{Supervision}$ *Employee*, whose notation is Ψ_1,
- Followee-Subordinate: *Employee* $\xleftarrow{Social^{-1}}$ *Employee* $\xleftarrow{Supervision^{-1}}$ *Employee*, whose notation is Ψ_2,
- Manager-Followee: *Employee* $\xleftarrow{Supervision}$ *Employee* $\xleftarrow{Social^{-1}}$ *Employee*, whose notation is Ψ_3,
- Subordinate-Followee: *Employee* $\xleftarrow{Supervision^{-1}}$ *Employee* $\xleftarrow{Social^{-1}}$ *Employee*, whose notation is Ψ_4,
- Followee-Peer: *Employee* $\xleftarrow{Social^{-1}}$ *Employee* $\xleftarrow{Supervision}$ *Employee* $\xleftarrow{Supervision^{-1}}$ *Employee*, whose notation is Ψ_5,
- Peer-Followee: *Employee* $\xleftarrow{Supervision}$ *Employee* $\xleftarrow{Supervision^{-1}}$ *Employee* $\xleftarrow{Social^{-1}}$ *Employee*, whose notation is Ψ_6,

where meta path, e.g., Ψ_1, denotes that the target employee receives information from his followee in online ESN, who gets information from his manager in the offline workplace. We can get $\mathcal{C}^{(hyb)} = \{\Psi_1, \Psi_2, \Psi_3, \Psi_4, \Psi_5, \Psi_6\}$ and $k^{(hyb)} = 6$. Based on each hybrid diffusion channel, e.g., Ψ_i, the amount of information on topic t that v sends to u can be represented as

$$w^{(hyb),i}(v,u,t) = \frac{2\left|\mathcal{P}^{(hyb)}_{\Psi_i}(v \rightsquigarrow u)\right| \cdot I(v,t)}{\left|\mathcal{P}^{(hyb)}_{\Psi_i}(v \rightsquigarrow \cdot)\right| + \left|\mathcal{P}^{(hyb)}_{\Psi_i}(\cdot \rightsquigarrow u)\right|}.$$

3.5 Channel Aggregation

Different diffusion channels deliver various amounts of information among employees via the online communications in ESN and offline contacts. In this subsection, we will focus on aggregating information propagated via different channels with the information aggregation function $f(\cdot) : \mathbb{R}^{n\times 1} \to [0,1]$, which can map the amount of information received by employees to their activation probabilities. Generally, any function that can map real number to probabilities in range $[0,1]$ can be applied and without loss of generality, we will use the logistic function $f(x) = \frac{e^x}{1+e^x}$ [4] in this paper.

Based on the information on topic t received by u via the online, offline and hybrid diffusion channels, we can represent u's activation probability to be:

$$f\left(\mathbf{w}^{(on)}(\cdot,u,t), \mathbf{w}^{(off)}(\cdot,u,t), \mathbf{w}^{(hyb)}(\cdot,u,t)\right) = \frac{e^{\left(g(\mathbf{w}^{(on)}(\cdot,u,t)) + g(\mathbf{w}^{(off)}(\cdot,u,t)) + g(\mathbf{w}^{(hyb)}(\cdot,u,t)) + \theta_0\right)}}{1 + e^{\left(g(\mathbf{w}^{(on)}(\cdot,u,t)) + g(\mathbf{w}^{(off)}(\cdot,u,t)) + g(\mathbf{w}^{(hyb)}(\cdot,u,t)) + \theta_0\right)}},$$

where function $g(\cdot)$ linearly combines the information in different channels belonging to certain sources and θ_0 denotes the weight of the constant factor. Terms $g(\mathbf{w}^{(on)}(\cdot,u,t))$, $g(\mathbf{w}^{(off)}(\cdot,u,t))$ and $g(\mathbf{w}^{(hyb)}(\cdot,u,t))$ can be represented as follows

$$g(\mathbf{w}^{(on)}(\cdot,u,t)) = \sum_{i=1}^{k^{(on)}} \alpha_i \cdot \sum_{v \in \Gamma^{(on),i}_{out}(u)} w^{(on),i}(v,u,t),$$

$$g(\mathbf{w}^{(off)}(\cdot,u,t)) = \sum_{i=1}^{k^{(off)}} \beta_i \cdot \sum_{v \in \Gamma^{(off),i}_{out}(u)} w^{(off),i}(v,u,t),$$

$$g(\mathbf{w}^{(hyb)}(\cdot,u,t)) = \sum_{i=1}^{k^{(hyb)}} \gamma_i \cdot \sum_{v \in \Gamma^{(hyb),i}_{out}(u)} w^{(hyb),i}(v,u,t),$$

where α_i, β_i, γ_i are the weights of different *online*, *offline* and *hybrid* diffusion channels respectively and $\sum_{i=1}^{k^{(on)}} \alpha_i + \sum_{i=1}^{k^{(off)}} \beta_i + \sum_{i=1}^{k^{(hyb)}} \gamma_i + \theta_0 = 1$. Depending of roles of different diffusion channels, the weights can be

- > 0, if positive information in the channel will increase employees' activation probability;
- $= 0$, if positive information in the channel will not change employees' activation probability;
- < 0, if positive information in the channel will decrease employees' activation probability.

In MUSE, weights of certain diffusion channels can be negative. As a result, the likelihood for a node to become active will no longer grow monotonically in the MUSE diffusion model. The optimal weights of different diffusion channels can be learned from the group participation log data (i.e., the target social activity diffusing at workplace). Different diffusion channels will be ranked according to their importance and top-k diffusion channels which can increase individuals' activation probabilities will be selected in the next subsection.

3.6 Channel Weighting and Selection

In Yammer, users can create and join groups of their interests, which can be about very diverse topics, e.g., products (e.g., iPhone, Windows, Android, etc.), people (e.g., Bill Gates, Leslie Lamport, etc.), projects (e.g., Project Complete, Meeting, ect.) and personal life issues (e.g., Diablo Games, Work Life Balance, etc.). The users' participation in groups log data can be represented as a set of tuples $\{(u,t)\}_{u,t}$, where tuple (u,t) represents that user u gets activated by topic t (of groups). Such a tuple set can be split into three parts according to ratio 3:1:1 in the order of the timestamps, where 3 folds are used as the training set, 1 fold is used as the validation set and 1 fold as the test set.

We will use the training set data to calculate the activation probabilities of individuals getting activated by topics in both the validation set and test set, while validation set is used to learn the weights of different diffusion channels and test set is used to evaluate the learned model.

Let $\mathcal{V} = \{(u,t)\}_{u,t}$ be the validation set. Based on the amount of information propagating among employees in the workplace calculated with the training set, we can infer the probability of user u's (who has not been activated yet) get activated by topic t, for $\forall(u,t) \in \mathcal{V}$, which can be represented with matrix $\mathbf{F} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{T}|}$, where $\mathbf{F}(i,j)$ denotes the inferred activation probability of tuple (u_i, t_j) in the validation set. Meanwhile, based on the validation set itself, we can get the ground-truth of users' group participation activities, which can be represented as a binary matrix $\mathbf{H} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{T}|}$. In matrix $\mathbf{H}$, only entries corresponding tuples in the validation set are filed with value 1 and the remaining entries are all filled with 0. The optimal weights of information delivered in different diffusion channels (i.e., α^*, β^*, γ^*, θ_0^*) can be obtained by solving the following objective function

$$\alpha^*, \beta^*, \gamma^*, \theta_0^* = \arg\min_{\alpha,\beta,\gamma,\theta_0} \|\mathbf{F} - \mathbf{H}\|_F^2$$
$$s.t. \sum_{i=1}^{k^{(on)}} \alpha_i + \sum_{i=1}^{k^{(off)}} \beta_i + \sum_{i=1}^{k^{(hyb)}} \gamma_i + \theta_0 = 1.$$

The final objective function is not convex and can have multiple local optima, as the aggregation function (i.e., the logistic function) is not convex actually. We propose to solve the objective function and handle the non-convex issue by using a two-stage process to ensure the robust of the learning process as much as possible.

(1) Firstly, the above objective function can be solved by using the method of Lagrange multipliers [2], where the corresponding Lagrangian function of the objective function can be represented as

$$\begin{aligned}
&\mathcal{L}(\alpha,\beta,\gamma,\theta_0,\eta)\\
&= \|\mathbf{F}-\mathbf{H}\|_F^2 + \eta(\sum_{i=1}^{k^{(on)}} \alpha_i + \sum_{i=1}^{k^{(off)}} \beta_i + \sum_{i=1}^{k^{(hyb)}} \gamma_i + \theta_0 - 1),\\
&= \mathrm{Tr}(\mathbf{F}\mathbf{F}^\top - \mathbf{F}\mathbf{H}^\top - \mathbf{H}\mathbf{F}^\top + \mathbf{H}\mathbf{H}^\top) + \eta(\sum_{i=1}^{k^{(on)}} \alpha_i + \sum_{i=1}^{k^{(off)}} \beta_i\\
&+ \sum_{i=1}^{k^{(hyb)}} \gamma_i + \theta_0 - 1).
\end{aligned}$$

By taking the partial derivatives of the Lagrange function with regards to variable $\alpha_i, i \in \{1,2,\cdots,k^{(on)}\}$, we can get

$$\begin{aligned}
\frac{\partial\mathcal{L}(\alpha,\beta,\gamma,\theta_0,\eta)}{\partial\alpha_i} &= \frac{\partial\mathrm{Tr}(\mathbf{F}\mathbf{F}^\top)}{\partial\alpha_i} - \frac{\partial\mathrm{Tr}(\mathbf{F}\mathbf{H}^\top)}{\partial\alpha_i} - \frac{\partial\mathrm{Tr}(\mathbf{H}\mathbf{F}^\top)}{\partial\alpha_i}\\
&+ \frac{\partial\mathrm{Tr}(\mathbf{H}\mathbf{H}^\top)}{\partial\alpha_i} + \frac{\partial\eta(\sum_{i=1}^{k^{(on)}}\alpha_i + \sum_{i=1}^{k^{(off)}}\beta_i + \sum_{i=1}^{k^{(hyb)}}\gamma_i + \theta_0 - 1)}{\partial\alpha_i}.
\end{aligned}$$

Term

$$\frac{\partial\eta(\sum_{i=1}^{k^{(on)}}\alpha_i + \sum_{i=1}^{k^{(off)}}\beta_i + \sum_{i=1}^{k^{(hyb)}}\gamma_i + \theta_0 - 1)}{\partial\alpha_i} = \eta$$

$$\frac{\partial\mathrm{Tr}(\mathbf{F}\mathbf{F}^\top)}{\partial\alpha_i} = \sum_{j=1}^{|\mathcal{U}|}\sum_{l=1}^{|\mathcal{T}|}\frac{\partial\mathbf{F}^2(j,l)}{\partial\alpha_i} = \sum_{j=1}^{|\mathcal{U}|}\sum_{l=1}^{|\mathcal{T}|} 2f(\mathbf{w}^{(on)}(\cdot,u_j,t_l),$$
$$\mathbf{w}^{(off)}(\cdot,u_j,t_l),\mathbf{w}^{(hyb)}(\cdot,u_j,t_l)) \cdot \frac{e^y}{(1+e^y)^2}\cdot\frac{\partial y}{\partial\alpha_i},$$

where the introduced term y denotes $y = g(\mathbf{w}^{(on)}(\cdot,u_j,t_l)) + g(\mathbf{w}^{(off)}(\cdot,u_j,t_l)) + g(\mathbf{w}^{(hyb)}(\cdot,u_j,t_l)) + \theta_0$ and its derivative is $\frac{\partial y}{\partial\alpha_i} = \frac{\partial g(\mathbf{w}^{(on)}(\cdot,u_j,t_l))}{\partial\alpha_i} = \sum_{v\in\Gamma_{out}^{(on),i}(u)} w^{(on),i}(v,u_j,t_k)$. Similarly, we can obtain terms $\frac{\partial\mathrm{Tr}(\mathbf{F}\mathbf{H}^\top)}{\partial\alpha_i}$, $\frac{\partial\mathrm{Tr}(\mathbf{H}\mathbf{F}^\top)}{\partial\alpha_i}$, and $\frac{\partial\mathrm{Tr}(\mathbf{H}\mathbf{H}^\top)}{\partial\alpha_i}$. By making $\frac{\partial\mathcal{L}(\alpha,\beta,\gamma,\theta_0,\eta)}{\partial\alpha_i} = 0$, we can obtain an equation involving variables α_i, β_i, γ_i, θ_0 and η. Furthermore, we can calculate the partial derivatives of the Lagrange function with regards to variable β_i, γ_i, θ_0 and η respectively and make the equation equal to 0, which will lead to an equation group about variables α_i, β_i, γ_i, θ_0 and η. The equation group can be solved with open source toolkits, e.g., SciPy Nonlinear Solver[2], effectively. By giving the variables with different initial values, multiple solutions (i.e., multiple local optimal points) can be obtained by resolving the objective function.

(2) Secondly, the local optimal points obtained are further applied to the objective function and the one achieving the lowest objective function value is selected as the final results (i.e., the weights of different channels).

According to the learned weights, different diffusion channels can be ranked according to their importance in delivering information to activate employees in the workplace. Considering that, some diffusion channels may not perform very well in information propagation (e.g., those with negative or zero learned weights), top-k channels that can increase employees' activation probabilities are selected as the effective channels used in MUSE model finally. In other words, k equals to the number of diffusion channels with positive weights learnt from the above objective function. Such a process is formally called diffusion channel weighting and selection in this paper.

The rational of channel weighting and selection is that: among all the diffusion channels, some channels can be useful but some may be not. 3 different sets of diffusion channels are introduced in previous sections and we want to select the good ones, which is quite similar to feature selection (selecting good features is helpful for improving the final prediction results). In the next section, we will show that channel weighting and selection can improve the performance of M^3USE a lot.

4. EXPERIMENTS

To examine the performance of MUSE in addressing the IDE problem, we conduct extensive experiments on real-world online ESN and organizational chart dataset in this section. Next, we will introduce the dataset used in the experiments briefly, give detailed descriptions about the experiment setting and analyze the experiment results.

4.1 Dataset Descriptions

We crawl all the Microsoft employees' information from Yammer and obtain the complete organizational chart involving all these employees in Microsoft during June, 2014

[2]http://docs.scipy.org/doc/scipy-0.14.0/reference/optimize.nonlin.html

[22, 21]. The social network data covers all the user-generated content (such as posts, replies, topics, etc.) and social graphs (such as user-user following links, user-group memberships, user-topic following links, etc.) by then that are set to be public. In summary, it includes more than $100k$ Microsoft employees, and millions of user-generated posts published and the social links.[3]

All the users in yammer are registered with the official employment ID in Microsoft, via which we can identify them in the organizational chart correspondingly. From Microsoft, the complete organization structure of all employees is obtained. As introduced before, the structure of the organizational chart is a rooted tree with the CEO at the top.

4.2 Experiment Settings

In the experiments, different groups represents different topics/themes and the employees' participations in certain groups represent that they are activated by the corresponding topics. Considering that some groups contain few members, which can pose great challenges in inferring the activation probabilities of employees getting activated by them, top 50 groups with the most members are selected from the dataset. All the employees who have participated in these 50 groups are organized into a set of (employee, group) tuples, which are split into 3 portions according to ratio $3:1:1$ (3 folds are used as the training set, 1 fold as the validation set and 1 fold as the test set). Employees who have been activated by groups in the training set will deliver information to other employees in the company about these groups via either online, offline or hybrid channels to activate them. To show that the proposed model **Muse** is a general diffusion model and can be applied to companies of different degrees of newness (i.e., sparsity of the known activation information), a subset of tuples are randomly sampled from the training set controlled by the parameter $\rho \in \{0.1, 0.2, \cdots, 0.9, 1.0\}$. We calculate the activation probabilities of all the (employees, group) tuples in the validation set, which together with the ground truth will be used to learn the weights of different diffusion channels. Top-k diffusion channels with the highest weights are selected, which will be used to build a new **Muse** model to infer the activation probabilities of (employee, group) tuples in the test set.

Comparison Methods To demonstrate the effectiveness of the proposed model, **Muse**, in this paper, we compare **Muse** with both state-of-art and traditional baseline diffusion models, which include:

- *Channel Weighting and Selection*: The diffusion model, **Muse**, proposed in this paper can extract online, offline and hybrid diffusion channels based on sets of social meta paths, whose weights can be learned from the activation log data. All the channels are ranked according to their weights and top-k channels with the largest weights are selected finally to build a new diffusion model **Muse** (new weights of the selected channels are learned).

- *Channel Weighting*: To verify the effectiveness of the channel selection step, we compare **Muse** with another diffusion model **Muse-w** in the experiments, where **Muse-w** is identical to **Muse** except that **Muse-w** uses all the proposed diffusion channels (without selection) and the weights of these channels are learned from the log data.

- *Channel with Fixed Weights*: Weight learning step can help **Muse** adjust the importance of information in different channels and fit the data better. To investigate such a statement, we also compare **Muse** with **Muse-fw** whose weights are fixed, i.e., $[\frac{1}{16}, \cdots, \frac{1}{16}]$, where $k^{(on)} + k^{(off)} + k^{(hyb)} = 16$.

- *Online ESN Only*: Offline organizational chart provides very important clues to infer potential offline contacts among employees. To show the impacts of the organizational structure, we further compare **Muse** with a state-of-art heterogeneous information diffusion model MLTM [6], which diffuses information within online ESN via multiple diffusion channels.

- *Offline Organizational Structure Only (implied by the organizational chart)*: Online ESN provides employees with very useful communication channels and to examine the effective of these channels, we extend the organization-based diffusion model CCM (Complex Cascade Model) proposed in [3] to the multi-topic case and compare it with **Muse** in the experiments.

- *Lossless ESN and Organizational Structure Coupling*: Diffusion model LCM (Lossless Coupling Model based on LT Model) proposed in [13] simply merges multiple homogeneous social networks into one network. We extend it to the online ESN (consisting of users and social links only) and offline organizational structure case and compare it with **Muse**.

- *Traditional LT Model*: For completeness, we also compare **Muse** with traditional linear threshold model LT merely based on the social links among employees in online ESN [8], where users' thresholds as well as the information diffusion weights of social links in LT is identical to those of **Muse**.

Evaluation Metrics

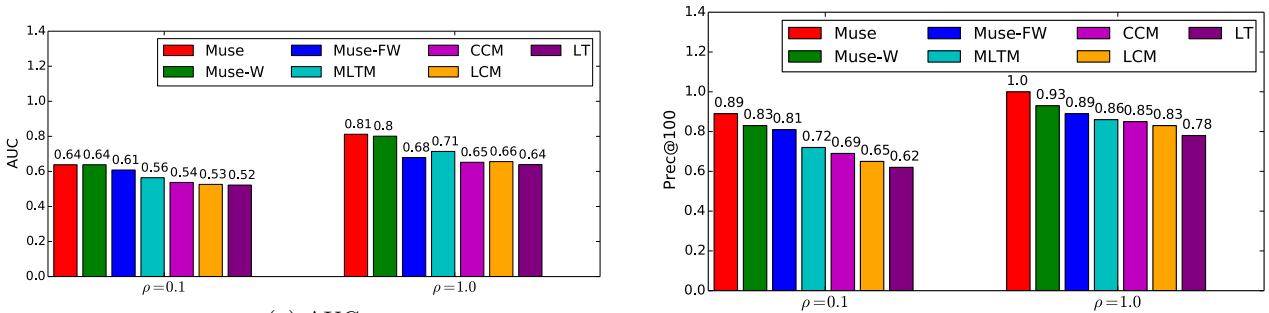
All these comparison methods can output the activation probabilities of users by various topics in the test set. To evaluate the performance of different comparison methods, two different metrics are applied in the experiments, which are AUC and Precision@100.

4.3 Experiment Results

The experiment results achieved by different comparison methods at the sampling rate $\rho = 0.1$ and $\rho = 1.0$ are available in Figure 2, which are evaluated by the AUC and Precision@100 metrics.

According to the results, we can observe that the method **Muse** can outperform the comparison methods **Muse-w** and **Muse-fw** when evaluated by both AUC and Precision@100. For instance, when $\rho = 1.0$ (i.e., all the training tuples are sampled), **Muse** achieves the AUC score of 0.81 and the Precision of 1.0, which are both higher than those obtained by **Muse-w** (i.e., AUC: 0.80 and Precision: 0.93), **Muse-fw** (AUC: 0.68, and Precision: 0.89). It demonstrates that diffusion channel selection step can improve the performance of **Muse** effectively.

[3]We are not able to reveal the actual numbers here and throughout the paper for commercial reasons.



(a) AUC

(b) Precision@100

Figure 2: Experiment results of comparison methods evaluated by AUC and Precision@100.

Table 2: Performance comparison of different diffusion models.

measure	methods	Activation tuple sampling ratio ρ. 0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AUC	**MUSE**	**0.692**	**0.728**	**0.750**	**0.767**	**0.781**	**0.791**	**0.799**	**0.805**
	MUSE-W	0.675	0.703	0.743	0.749	0.768	0.780	0.789	0.794
	MUSE-FW	0.618	0.627	0.640	0.648	0.665	0.672	0.674	0.674
	MLTM	0.567	0.574	0.586	0.601	0.617	0.644	0.686	0.708
	CCM	0.563	0.570	0.585	0.596	0.618	0.620	0.648	0.649
	LCM	0.547	0.570	0.585	0.601	0.616	0.627	0.638	0.647
	LT	0.541	0.560	0.574	0.588	0.603	0.612	0.621	0.631
Precision@100	**MUSE**	**0.91**	0.91	**0.95**	**0.95**	**0.98**	**1.0**	**1.0**	**1.0**
	MUSE-W	0.86	0.92	**0.92**	0.93	0.93	0.92	0.93	0.93
	MUSE-FW	0.81	0.85	0.86	0.86	0.87	0.86	0.87	0.87
	MLTM	0.75	0.77	0.79	0.81	0.81	0.83	0.83	0.86
	CCM	0.69	0.69	0.76	0.76	0.76	0.78	0.79	0.79
	LCM	0.68	0.68	0.74	0.73	0.75	0.75	0.75	0.82
	LT	0.65	0.68	0.72	0.74	0.74	0.75	0.75	0.76

Diffusion channels weighted by learned importance from the data can better model the diffusion process in the real world. To support such a claim, we compare **MUSE-W** with **MUSE-FW**. According to the results, we observe that **MUSE-W** achieves much better performance than **MUSE-FW** (with fixed weights). For example, when the sampling ratio $\rho = 1.0$, the AUC obtained by **MUSE-W** is 0.80, which is over 17% larger than the AUC obtained by **MUSE-FW**; the Precision@100 achieved by **MUSE-W** is 0.93, which outperform **MUSE-FW** by over 5%.

We also compare **MUSE** with state-of-art single source diffusion models, which include diffusion model designed for multi-relational networks (i.e., MLTM) and that proposed for organizational structure only (i.e., CCM). Compared with MLTM and CCM, we observe that **MUSE** which utilize the connection information across online ESN and offline organizational structure simultaneously can achieve far better performance. For example, at $\rho = 0.1$, the AUC score achieved by **MUSE** is 0.64, which is over 18% higher than that obtained by MLTM, CCM, LCM and LT. Meanwhile, the Precision@100 score obtained by **MUSE** at $\rho = 0.1$ is 0.89, which is over 23% greater than that achieved by MLTM, 29% greater than that of CCM and more than 35% larger than that obtained by LCM and LT.

Heterogeneous information (besides the social links and supervision links) in both online ESN and offline organizational structure creates more diverse connections among employees, which can be demonstrated by comparing **MUSE** with LCM, which is a extension of the state-of-art diffusion model proposed for multi-homogeneous social networks. The advantages of **MUSE** over LCM is very obvious according to the result in Table 2. For completeness, we also compare **MUSE** with the traditional LT model merely based on the social links among employees at workplace and **MUSE** can out-perform LT with significant advantages.

The performance of these methods can be affected by the sampling ratio parameter ρ a lot. Therefore, we have also conducted the parameter analysis about ρ, and the results are shown in Table 2. As shown in the table, we change ρ with values in $\{0.2, 0.3, \cdots, 0.9\}$, and evaluate the comparison methods' performance by metrics AUC and Precision@100. Generally, the performance of all the comparison methods improves steadily the sampling rate ρ increases, and **MUSE** can outperform the other baseline methods with great advantages consistently.

4.4 Channel Importance Analysis

Various diffusion channels have been extracted across the online ESN and offline organizational structure, which will be weighted according to their importance in Section 3.6. For all the channels, we obtain their weights learned from the data in model **MUSE-W** (all the channels without selection) when $\rho = 1.0$, which are ranked and listed in Table 3.

In Table 3, we display the rank, notation, and physical meanings of all the 16 online, offline and hybrid diffusion

Table 3: Rank of different diffusion channels.

Rank	Channal Notation	Channal Physical Meaning
1	Ω_1	"Manager"
2	Φ_1	"Followee"
3	Ω_4	"2nd-Level Manager"
4	Ψ_6	"Peer-Followee"
5	Ψ_3	"Manager-Followee"
6	Ω_3	"Peer"
7	Φ_2	"Followee-Followee"
8	Ψ_1	"Followee-Manager"
9	Ψ_5	"Followee-Peer"
10	Ψ_4	"Subordinate-Followee"
11	Ψ_2	"Followee-Subordinate"
12	Ω_2	"Subordinate"
13	Ω_5	"2nd-Level Subordinate"
14	Φ_3	"Reply Post"
15	Φ_5	"Post Notification"
16	Φ_4	"Like Post"

channels extracted in this paper. We observe that the diffusion channel with the largest weight is Ω_1 (Manager), which represents that information delivered from managers to subordinates plays a more important role than other channels. The 2_{nd} important diffusion channel is Φ_1 (Followee), i.e., the social links among employees in the online ESN, which denotes that employees' group participation activities has very high correlation with the social links among them online. Diffusion channels of ranks 3 to 5 are Ω_4 (2nd-Level Manager), Ψ_6 (Peer-Followee) and Ψ_3 (Manager-Followee) respectively, two of which are the hybrid diffusion channels across the online ESN and offline organizational structure. It supports our motivation of extracting hybrid diffusion channels at Section 3.4.

At the bottom of Table 3, we observe that diffusion channels with the lowest weights are Ω_2 (Subordinate), Ω_5 (2nd-Level Subordinate), Φ_3 (Reply Post), Φ_5 (Post Notification) and Φ_4 (Like Post). Channels Ω_2 and Ω_5 being ranked at the end of the list shows that information delivered from subordinates to managers has little effects on the group participation activities of managers. Meanwhile, the diffusion channels associated with posts are ranked at the last shows that, in online ESN, little group participation activity information diffuses via posts compared with other online diffusion channels.

By comparing all the online ESN diffusion channels, we see that diffusion channels consisting of social links only (including both Φ_1 and Φ_2) are more important than those involving posts (i.e., Φ_3, Φ_4 and Φ_5). In addition, compared with Φ_2 (Followee-Followee), channel Φ_1 (Followee) of shorter length is more effective, i.e., employees directed connected have larger impact on each other than those who are not directly connected.

By comparing all the inferred offline diffusion channels, we observe that channels from managers (Ω_1 and Ω_4), from peers (Ω_3) and those from subordinates Ω_2 and Ω_5 have relation: "Manager > Peer > Subordinate", i.e., managers have greater impacts on subordinates than that among peers than that from subordinates on managers.

By comparing all the hybrid diffusion channels, we find that they are all effective channels with ranks from 4 to 11 and employees tend to trust their managers, peers more than their subordinates, as Ψ_6, Ψ_3, Ψ_1 and Ψ_5 are ranked ahead of Ψ_4 and Ψ_2.

In addition, generally speaking, diffusion channels of shorter length (e.g., Ω_1, Φ_1 except Ω_2) are more effective than longer channels in diffusing information in workplace (e.g., Ω_3, Ω_4, Ψ_3 and Ψ_6, etc.).

5. RELATED WORK

Information diffusion has been a hot research topic in the last decade and dozens of papers have been published on this topic so far. Domingos and Richardson [5, 14] are the first to propose to study the influence propagation based on knowledge-sharing sites. Kempe et al. [8] are the first ones to study the influence propagation problem through social networks and propose two famous diffusion models: Independent Cascade (IC) model and Linear Threshold (LT) model, which have been the basis of many diffusion models proposed later.

Based on the diffusion models, various application can be studied. For example, Kempe et al. [8] also study the seed user selection problem to maximize the influence within the social network (i.e., influence maximization problem or viral marketing problem). A large-scale network social influence analysis model is introduced by Tang et al. in [17], which is implemented and under the Map-Reduce framework. Gui et al. [6] models the diffusion of research topics among researchers in the bibliographic network, while Chelmis et al. propose to study the role of organizational chart in diffusing information from managers to subordinates in a company.

In recent years, studying multiple networks simultaneously has become a hot research topic. Kong et al. notice that people are usually involved in multiple social networks and propose to identify the correspondence relationships between the shared users in two fully aligned social networks [10]. Based on the aligned networks, Zhang et al. [23] propose to transfer link across different social networks to study the link recommendation problems. A complete survey about the cross-network research problems and recent works done in heterogeneous information networks is available in [15].

Meanwhile, some works have also been done on studying information diffusion problems by considering multiple networks/sources. Zhan et al. [19] propose to study the influence maximization problem across two partially aligned social networks based on the extracted multi-relations among users. In online social networks, sometimes a large number of group members may rapidly and dramatically change their behavior by widely adopting a previously rare practice, the time point of which is defined as the *tipping point* in [20]. Zhan et al. propose to identify the *tipping users* whose involvement can effectively trigger the *tipping point* across multiple aligned social networks [20]. Nguyen et al. [13] propose a coupling-based diffusion models to study the influence maximization problem in multiplex social networks. Myers et al. [12] and Lin et al. [11] present two different information diffusion models incorporating both external influence sources as well as the internal influence among users in online social networks.

6. CONCLUSION

In this paper, we have studied the IDE problem about information diffusion among employees at workplace. To solve the IDE problem, a novel diffusion model **Muse** is introduced. Based on the heterogeneous information in on-

line ESNs and offline organizational chart, 3 sets of different diffusion channels have been extracted by MUSE, information diffused via which is weighted and aggregated in MUSE. The optimal weights of different diffusion channels are learned from the target social activity log data in MUSE and top-k useful diffusion channels are selected finally. Experiments conducted on real-world online ESN and organizational chart demonstrate the effectiveness of MUSE.

7. ACKNOWLEDGEMENT

This work is supported in part by NSF through grants III-1526499.

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

8. REFERENCES

[1] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *WWW*, 2012.

[2] D. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series)*. Athena Scientific, 1996.

[3] C. Chelmis and V. Prasanna. The role of organization hierarchy in technology adoption at the workplace. In *ASONAM*, 2013.

[4] John S. deCani and Robert A. Stine. A note on deriving the information matrix for a logistic distribution. *The American Statistician*, 1986.

[5] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, 2001.

[6] H. Gui, Y. Sun, J. Han, and G. Brova. Modeling topic diffusion in multi-relational bibliographic information networks. In *CIKM*, 2014.

[7] E. Keeler. The value of remaining lifetime is close to estimated values of life. *Journal of Health Economics*, 2000.

[8] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[9] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *STOC*, 2000.

[10] X. Kong, J. Zhang, and P. Yu. Inferring anchor links across multiple heterogeneous social networks. In *CIKM*, 2013.

[11] S. Lin, F. Wang, Q. Hu, and P. Yu. Extracting social events for learning better information diffusion models. In *KDD*, 2013.

[12] S. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *KDD*, 2012.

[13] D. Nguyen, H. Zhang, S. Das, M. Thai, and T. Dinh. Least cost influence in multiplex social networks: Model representation and analysis. In *ICDM*, 2013.

[14] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, 2002.

[15] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. Yu. A survey of heterogeneous information network analysis. *CoRR*, abs/1511.04854, 2015.

[16] Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *VLDB*, 2011.

[17] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD*, 2009.

[18] T. Turner, P. Qvarfordt, J. Biehl, G. Golovchinsky, and M. Back. Exploring the workplace communication ecology. In *CHI*, 2010.

[19] Q. Zhan, J. Zhang, S. Wang, P. Yu, and J. Xie. Influence maximization across partially aligned heterogenous social networks. In *PAKDD*, 2015.

[20] Q. Zhan, J. Zhang, P. Yu, S. Emery, and J. Xie. Discover tipping users for cross network influencing. In *IRI*, 2016.

[21] J. Zhang, Y. Lv, and P. Yu. Enterprise social link recommendation. In *CIKM*, 2015.

[22] J. Zhang, P. Yu, and Y. Lv. Organizational chart inference. In *KDD*, 2015.

[23] J. Zhang, P. Yu, and Zhou Z. Meta-path based multi-network collective link prediction. In *KDD*, 2014.