

Measuring Friendship Closeness: A Perspective of Social Identity Theory

Shiqi Zhang*
National University of Singapore
Singapore, Singapore
Southern University of Science and Technology
Shenzhen, China
s-zhang@comp.nus.edu.sg

Xiaokui Xiao
National University of Singapore
Singapore, Singapore
xkxiao@nus.edu.sg

Jiachen Sun
Wenqing Lin[†]
jiachensun@tencent.com
edwlin@tencent.com
Tencent Inc.
Shenzhen, China

Bo Tang[†]
Southern University of Science and Technology
Shenzhen, China
tangb3@sustech.edu.cn

ABSTRACT

Measuring the closeness of friendships is an important problem that finds numerous applications in practice. For example, online gaming platforms often host friendship-enhancing events in which a user (called the source) only invites his/her friend (called the target) to play together. In this scenario, the measure of friendship closeness is the backbone for understanding source invitation and target adoption behaviors, and underpins the recommendation of promising targets for the sources. However, most existing measures for friendship closeness only consider the information between the source and target but ignore the information of groups where they are located, which renders inferior results. To address this issue, we present new measures for friendship closeness based on the social identity theory (SIT), which describes the inclination that a target endorses behaviors of users inside the same group. The core of SIT is the process that a target assesses groups of users as them or us. Unfortunately, this process is difficult to be captured due to perceptual factors. To this end, we seamlessly reify the factors of SIT into quantitative measures, which consider local and global information of a target's group. We conduct extensive experiments to evaluate the effectiveness of our proposal against 8 state-of-the-art methods on 3 online gaming datasets. In particular, we demonstrate that our solution can outperform the best competitor on the behavior prediction (resp. online target recommendation) by up to 23.2% (resp. 34.2%) in the corresponding evaluation metric.

CCS CONCEPTS

• **Information systems** → **Massively multiplayer online games; Social networks**; *Social recommendation*.

*This work was done while Shiqi Zhang was an intern at Tencent.

[†]Wenqing Lin and Bo Tang are the corresponding authors.



This work is licensed under a Creative Commons Attribution International 4.0 License.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9236-5/22/10.
<https://doi.org/10.1145/3511808.3557076>

KEYWORDS

Graph; Closeness Measure; Social Science

ACM Reference Format:

Shiqi Zhang, Jiachen Sun, Wenqing Lin, Xiaokui Xiao, and Bo Tang. 2022. Measuring Friendship Closeness: A Perspective of Social Identity Theory. In *Proceedings of the 31st ACM Int'l Conference on Information and Knowledge Management (CIKM '22)*, Oct. 17–21, 2022, Atlanta, GA, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3511808.3557076>

1 INTRODUCTION

Consider a social network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v_s \in \mathcal{V}$ represents the user v_s and each edge $(v_s, v_t) \in \mathcal{E}$ represents the friendship of two users v_s and v_t . Measuring the closeness of the friendship (v_s, v_t) is an important problem and finds numerous applications in real-world scenarios, where a user v_s (called the source) *only* interacts with the friend v_t (called the target). For example, online gaming platforms often host friendship-enhancing events, which encourage the source v_s to invite the target v_t to return to the game or play together [26, 27]. In this scenario, the friendship closeness measure is the backbone for understanding the invitation (resp. adoption) behavior of v_s (resp. v_t). Furthermore, this measure can be applied in recommending promising targets for the sources because the number of targets of a source could be considerable. Analogously, in the instant messaging platform WeChat, a source's post can only be liked and commented on by the target in contacts [56], thus the friendship closeness measure can also be leveraged to predict corresponding activities [17, 37, 55].

In the present work, we focus on the *topological friendship closeness* (TFC) measure that reflects the closeness of a source-target pair (v_s, v_t) on \mathcal{G} rather than labels associated with them. In particular, most existing TFC measures [1, 13, 18–20, 30, 31, 36, 48, 51, 54] adopt the structural information between two users v_s and v_t . For example, [13] directly employs the tie strength of (v_s, v_t) , which is measured by the times of historical interactions: [1] proposes several measures in terms of the common neighborhood of (v_s, v_t) . Furthermore, some measures preserve long and intricate paths between (v_s, v_t) based on random walks, e.g., personalized PageRank [30]. Unfortunately, the above-said measures may render compromised results, as they ignore the group information and the potential

group effect related to (v_s, v_t) , which says a user's belonging group can influence his/her decision.

Even though some TFC measures [2, 8, 11, 43, 45, 50] attempt to take the group effect into consideration, the involved information is still inadequate. To explain, we find that these measures typically require (i) first representing the affiliation relationship between users and groups as a bipartite graph, where users and groups are two disjoint sets of nodes and affiliation relationships are a set of edges, and (ii) then exploiting information on this bipartite graph. For example, [50] proposes structural diversity to measure the number of adjacent groups for a given user; [2, 8, 43] utilize edge weights between users and groups; [11, 45] measures the commonality of groups between two users on this bipartite graph. However, in these works, contracting a group into a node makes more detailed group information imperceptible, e.g., connectivity.

To mitigate the deficiencies of existing measures, we propose to explore the group effect by leveraging the social identity theory (SIT), which is a fundamental concept in social psychology and is widely applied in domains of human health [38], team sports [35], computer-supported cooperative works [39] and fake news detection [40]. SIT describes that a target v_t tends to endorse attitudes and behaviors of groups of users, who are assessed as us in v_t 's cognition. In other words, given a target v_t and an identified group C of v_t , v_t is more likely to adopt the invitation from the source $v_s \in C$. Unfortunately, the identification process is difficult to be quantified due to perceptual factors. To this end, we seamlessly reify the factors of SIT into TFC measures, which consider a group's local and global information. For example, by modeling the social network as a physical system, we propose to measure the attractive spring-like force related to a group, in which the information of group connectivity and in-group user homogeneity are naturally incorporated. In contrast, previous TFC measures only consider similarities of group memberships [11, 45].

We experimentally evaluate the proposed SIT-based solution against 8 representative competitors on 3 real-world online game datasets. In particular, we demonstrate that the proposed solution outperforms all competitors in terms of AUC, accuracy, and F1 score while predicting target adoption and source invitation behaviors on the tested dataset. Besides, we detailedly analyze the relations between user behaviors and SIT factors, and the importance of these factors. At last, we deploy our solution to the online target recommendation, which achieves up to 34.2% improvement over the best treatment in the corresponding evaluation metric.

To summarize, we make the following contributions in this work:

- We propose new TFC measures based on SIT in social psychology, which preserve both local and global information of groups.
- We conduct extensive experiments and analysis to demonstrate the superiority of proposed SIT-based TFC measures over state-of-the-art competitors while predicting user behaviors.
- We deploy the presented solution to the online target recommendation, which achieves significant improvement.

2 PRELIMINARIES

In this section, we first elaborate on the problem of measuring the topological friendship closeness (TFC), followed by the illustration of main competitors and downstream tasks that TFC measures are

to solve in this work. At last, we introduce the background of the social identity theory.

2.1 Problem Formulation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a social network, where \mathcal{V} is a set of users (called nodes) and \mathcal{E} is a set of friendships (called edges). We assume that the friendship $(v_s, v_t) \in \mathcal{E}$ is directed and is associated with an edge weight $w_{s,t} \in (0, 1]$. Given a directed edge $(v_s, v_t) \in \mathcal{E}$, we call v_s (resp. v_t) the *source* (resp. *target*) neighbor of v_t (resp. v_s). Given a node $v_t \in \mathcal{V}$, the *source neighborhood* of v_t consists of: (i) v_t itself; (ii) the node set \mathcal{N}_t , which contains sources of v_t ; (iii) the edge set \mathcal{E}_t , which contains the edge $(v_s, v_t) \in \mathcal{E}$; (iv) the edge set \mathcal{R}_t , which contains the edge $(v_s, v_k) \in \mathcal{E}$ between sources $v_s, v_k \in \mathcal{N}_t$.

Recall in Section 1 that, given an input social network \mathcal{G} , the TFC measure reflects the topological closeness for each $(v_s, v_t) \in \mathcal{E}$. Given a source-target pair $(v_s, v_t) \in \mathcal{E}$, the goal of our present work is to design the TFC measure, which leverages the structural information of the group C , where $v_s, v_t \in C$. For simplicity, we assume that v_s and v_t are only co-located in one group C . As for the partitioning of groups, it relies on the social identity theory and will be illustrated in Section 3.

2.2 Main Competitors

In what follows, we briefly elaborate on the main ideas of competing TFC measures in the present work. For better clarity, we call a measure as the *group-level* TFC measure if considering the group information, and the *individual-level* TFC measure otherwise. Specifically, we select 5 representative individual-level measures: tie strength [13, 48], number of common neighbors [1, 48], personalized PageRank [30], cosine and euclidean similarity of Node2Vec vectors [19]; 3 commonly-used group-level measures: structural diversity [50], user-group tie strength [43] and group edge density [32]. It is worth noting that the selected individual-level measures are also known as proximities [5, 48], which can reflect the topological closeness for all node pairs. Furthermore, these TFC measures are widely applied in Tencent gaming platforms [26, 27].

Tie strength. Given a friendship $(v_s, v_t) \in \mathcal{E}$, the tie strength equals to the edge weight $w_{s,t}$ [13, 48].

Number of common neighbors. Given a friendship $(v_s, v_t) \in \mathcal{E}$, this measure counts the number of common neighbors [1, 48] of source v_s and target v_t , which can reflect the similarity of two node's local structure.

Personalized PageRank. Given a friendship $(v_s, v_t) \in \mathcal{E}$, this measure aims to preserve intricate topological relations between nodes by performing random walks. In particular, the personalized PageRank [30] from v_s to v_t is defined as the probability that a *random walk with restart* (RWR) [49] originating from v_s stops at v_t . The corresponding RWR starts from the source v_s and, in each step, chooses to (i) either terminate at the current node with probability α , (ii) or navigate to a random target neighbor of the current node with the remaining $1 - \alpha$ probability.

Similarity of Node2Vec vectors. As a node embedding method [5], Node2Vec [19] proposes to represent each node $v_s \in \mathcal{G}$ as a compact vector, which preserves the structural information in the vicinity of v_s . By this method, given a friendship $(v_s, v_t) \in \mathcal{E}$, the TFC can be

measured by the cosine similarity or euclidean distance between the Node2Vec vectors of v_s and v_t . Compared with the prior work DeepWalk [31], the core contribution of Node2Vec is that it takes *two-order random walks* as inputs for training. Compared with the RWR, the two-order random walk will not terminate until a pre-defined length limit. Moreover, given the current node v_j , the target neighbors of v_j are first divided into the following three categories: (i) the source neighbor v_i that was visited in the previous step, (ii) v_k that is the common target neighbor of v_i or (iii) otherwise. Then different (resp. same) transition probabilities are assigned to nodes across (resp. inside) each category. At last, the next node of this walk is selected in terms of the biased probability.

Structural diversity. Ugander et al. are the first to propose the concept of structural diversity [50], which is further applied for influence analysis [13, 34, 46, 55]. Specifically, given a friendship $(v_s, v_t) \in \mathcal{E}$, the TFC measure based on the structural diversity equals to the number of weakly connected components (see Definition 2.1) in a subgraph, which is derived from v_t 's source neighbors affected by a specific event and their relationships. This work [50] points out that a more extensive structural diversity of v_t indicates a higher probability that v_t will be influenced.

Definition 2.1 (Weakly Connected Component). Given a directed graph, a weakly connected component is a maximal subgraph, where any inside node is connected by paths if ignoring the direction of edges.

User-group tie strength. Given a friendship $(v_s, v_t) \in \mathcal{E}$ and a group C where $v_s, v_t \in C$, the TFC measure based on the user-group tie strength [2, 8, 43] equals to the summation of tie strength between v_t and other members in C .

Group edge density. Given a friendship $(v_s, v_t) \in \mathcal{E}$ and a group C where $v_s, v_t \in C$, the TFC measure based on the group edge density is defined as the fraction of the summation of edge weights in C over the number of all possible edges in C , which is also applied to analyze group dynamics [32, 34].

2.3 Downstream Tasks

Before introducing downstream tasks, we first illustrate friendship-enhancing events in Tencent's online gaming platforms, whose procedure is as follows. First, given a social network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a source set $\mathcal{S} \subseteq \mathcal{V}$ and a target set $\mathcal{T} \subseteq \mathcal{V}$ are selected regarding the event demand, e.g., recalling inactive targets and stimulating interactions between friends. Then, the event is sent to each source $v_s \in \mathcal{S}$, and allows v_s to invite the target neighbor $v_t \in \mathcal{T}$ from a feed window with at most k target friends. Once a target v_t adopts the invitation from the source v_s , both v_s and v_t will receive virtual gifts as incentives. To summarize, there exist two types of user activities: *source invitation* and *target adoption*. In friendship-enhancing events, TFC measures act as structural heuristics to facilitate the understanding of invitation and adoption behaviors and recommending targets for the sources, which are defined as follows.

PROBLEM 1 (BEHAVIOR PREDICTION). *Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a source-target pair $(v_s, v_t) \in \mathcal{E}$, the objective of behavior prediction is to infer the likelihood that source invitation and target adoption behaviors happen between (v_s, v_t) .*

Table 1: Factors in SIT.

Factor	Meaning
Multi-membership	Number of groups
Inclusiveness	Number of in-group members
Satisfaction	v_t 's feelings about being a group member
Solidarity	v_t 's psychological bond with in-group members
Centrality	Importance of a group in v_t 's cognition
Self-stereotyping	Similarity of v_t and group average in v_t 's cognition
In-group homogeneity	Similarity within a group
Social standing	Social standing of a group

PROBLEM 2 (TARGET RECOMMENDATION). *Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a budget k , a source set \mathcal{S} and a target set \mathcal{T} , the objective of target recommendation is to select at most k target neighbors from \mathcal{T} for each source user $v_s \in \mathcal{S}$, such that the likelihood that both invitation and adoption behaviors happen among all returned source-target pairs are maximized.*

2.4 Social Identity Theory

Formulation. Turner and Tajfel define social identity as 'the individual's knowledge that he/she belongs to certain social groups together with some emotional and value significance to him of this group membership' [47]. In other words, social identity theory (SIT) emphasizes how the target user's belonging group affects his/her behavior. Given a social network G and a target v_t , SIT is formulated by three cognitive processes: *categorization*, *identification* and *comparison* [47]. Specifically, the categorization is a preprocessing step, which partitions the users of \mathcal{G} into several candidate social groups w.r.t v_t . In the identification stage, v_t will identify him/herself as a member of a certain candidate group. Finally, the comparison stage says that, for the sake of self-esteem, v_t tends to shift the sense of belonging based on group social standing.

Factors. After pre-partitioning candidate groups for a target v_t , both identification and comparison stages depend on many factors. We summarize the factors used in this paper in Table 1. Specifically, to understand the identification stage, the statistical factors multi-membership [41] and inclusiveness [3] are proposed. Besides, Leach et al. [21] extensively summarize existing social perceptual factors as: satisfaction [6, 47], solidarity [6, 9], centrality [3, 6, 9], self-stereotyping [29, 42] and in-group homogeneity [29, 42]. Regarding the comparison stage, by definition, the social standing of a group is also recognized as a pivotal factor [47].

3 OVERVIEW

We illustrate the main workflow of the present work in Figure 1, which processes the input graph by following three steps.

Step 1: categorization. For employing SIT, the first step is partitioning candidate groups for each target user, i.e., the categorization stage. To explore the group information for the source-target pair, given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a target user v_t , we partition the source neighborhood of v_t based on the weakly connected components (see Definition 2.1). Specifically, following operations in [11, 13, 50, 55], we first extract each weakly connected component (dubbed as CC) from the ego network $\mathcal{G}_t = (\mathcal{N}_t, \mathcal{R}_t)$ of v_t , and then call all sources in a given CC and the target v_t as a candidate

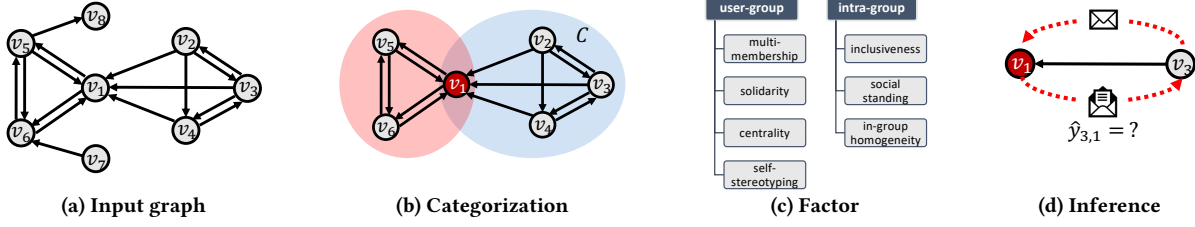


Figure 1: Illustration of core subroutines.

group C , since v_t connects to all sources in the CC by edges in \mathcal{E}_t . To exemplify, given an input graph as shown in Figure 1(a), the source neighborhood of target node v_1 is partitioned into two candidate groups (shaded in two colors in Figure 1(b)), among which the group C is constituted by v_1, v_2, v_3, v_4 . The rationale for employing CC is two-fold. First, as suggested in SIT [15], the candidate group requires to be categorized in a more general way to mitigate the in-group bias. For instance, employers in two departments of the same company should be categorized into one company group rather than split into two department groups. By definition, the CC can satisfy this requirement, as the inside nodes of two CCs are isolated (*i.e.*, no path). Second, CC is a simple but effective way to represent the community in the ego network of a user [11, 53]. As evidence, [11] finds that employing CCs for the link prediction can achieve comparable or even slightly better performance than employing communities detected by the state-of-the-art methods.

Step 2: SIT-based measure definition. As shown in Figure 1(c), factors of SIT in Table 1 can be separated into two branches: (i) user-group factors, describing the information between the target and other members; (ii) intra-group factors, describing the stand-alone information of other members. For example, the factors of self-stereotyping and in-group homogeneity both focus on individual similarities, however, by definition, the former is a user-group factor and the latter is an intra-group one. Accordingly, once the categorization stage is accomplished, given a target v_t and a candidate group C , the second step is to devise group-level measures between v_t and $C \setminus v_t$ (resp. among sources in $C \setminus v_t$) to quantify user-group (resp. intra-group) factors of SIT. Notice that we skip the satisfaction factor as it involves numerous human sentiments and is difficult to be detected by employing graph topology only.

Step 3: Inclination inference. In the third step, we employ the proposed SIT-based measures to infer the inclination that the target v_t endorses each candidate group C . This inclination is defined as

$$\hat{y}_{C,t} = \frac{\sum_{v_s \in C \setminus v_t} \hat{y}_{s,t}}{|C| - 1},$$

where $|C| \geq 2$ as there exist at least v_t and a source, and $\hat{y}_{s,t}$ is the likelihood that v_t adopts the invitation from each source v_s in C (see Figure 1(d)) and can be applied in the downstream Problem 1 and Problem 2. Without loss of generality, we derive $\hat{y}_{s,t}$ by a supervised manner, which works as follows. For each training source-target pair (v_s, v_t) , we first take the proposed SIT-based measures as d -dimensional features $x_{s,t} \in \mathbb{R}^d$ and $y_{s,t} \in \{0, 1\}$ as the label of the adoption behavior, where $y_{s,t} = 1$ indicates v_t adopts the invitation from v_s and $y_{s,t} = 0$ otherwise. We next train a well-accepted XGBoost [7] model, which is finally utilized to

infer each $\hat{y}_{s,t}$. Specifically, given a training dataset with related tuple $(x_{s,t}, y_{s,t})$ and T regression trees with the maximum depth h , XGBoost aims to find the best parameters attached on all possible 2^h leaves of each of T trees by the following objective where all parameters are represented by a matrix $\Theta \in \mathbb{R}^{T \times 2^h}$,

$$\arg \min_{\Theta} (\mathcal{L}(\Theta) + \Omega(\Theta)). \quad (1)$$

In Eq.(1), the training loss term is defined as the logistic loss

$$\mathcal{L}(\Theta) = \sum_{s,t} \left(y_{s,t} \ln(1 + e^{-\hat{y}_{s,t}}) + (1 - y_{s,t}) \ln(1 + e^{\hat{y}_{s,t}}) \right), \quad (2)$$

and the regularization term $\Omega(\Theta)$ considers both L0 and L2 norms of Θ . Regarding the predicted value $\hat{y}_{s,t}$ in Eq.(2), it aggregates the parameters by $\hat{y}_{s,t} = \sum_i \Theta[i, q_i(s, t)]$, where the i -th row of Θ contains all parameters in the i -th tree and $q_i(s, t)$ is the column index that contains the parameter of current (v_s, v_t) in the i -th tree. The reason of employing XGBoost is that (i) it is an empirically-efficient solver [7] and is widely accepted for massive game data in Tencent; (ii) the model is easy to be interpreted and is further applied for related analysis in Section 5.2.

Remarks. In what follows, we focus on defining SIT-based measures in Section 4, where our key contributions come from. The performance of two downstream tasks are evaluated in Section 5.2 and Section 5.3. As selections of community detection method in step 1 and training model in step 3 are orthogonal to our work, we refer the interested readers to [11] for the benchmark of different community detection methods and [7] for the detailed training procedure of XGBoost, respectively.

4 SIT-BASED MEASURES

In this part, we devise the formulation for SIT factors and illustrate their rationales compared with existing group-level measures.

4.1 Multi-membership and Inclusiveness

Inspired from the structural diversity [50], we utilize the number of candidate groups and the size of each group of a target to represent the statistical factors of multi-membership and inclusiveness, respectively. In particular, given a graph \mathcal{G} and a target user v_t , we define the multi-membership of v_t as the number of candidate groups (#CC) consisting of *all* source neighbors of v_t , and the inclusiveness of a given group C as the group size (GS), *i.e.*, the cardinality $|C|$. In contrast, the prior work of structural diversity [50] focuses on the groups with source neighbors who have sent the invitation only.

The rationales are as follows. First, before the release time of a given event, invitation behaviors of source users are unknown, and hence corresponding ego networks fail to be extracted. Second,

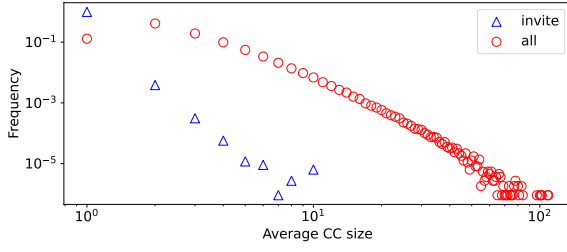


Figure 2: Distribution of averaged CC size in FPS.

even though the ego networks may be constructed by exploiting source behaviors from historical events, however, the invitation behaviors are only triggered by a small number of source users, which renders a much sparser ego network with many singleton CCs (*i.e.*, one inside member) [50] and makes less group information revealed. For example, Figure 2 shows the distributions of averaged CC size of invited target users on FPS dataset, in which averaged CC size of a target v_t equals the fraction of the number of all (or inviting) source neighbors over #CC of v_t . Specifically, by using the inviting source nodes, the resulting distribution in blue triangles is highly-skewed, where singletons are over 99%. In contrast, by using all source neighbors, the results in red circles show an improvement w.r.t the average size, which reveals more group information.

4.2 Social Standing and Centrality

Recall in Section 2.4 that factors of social standing and centrality in SIT describe the group importance in the whole graph and a given target’s cognition, which can be quantified by PageRank and personalized PageRank (PPR), respectively. Given a graph \mathcal{G} with $|\mathcal{V}| = n$ and a restart probability α , the PageRank vector of \mathcal{G} is denoted as a vector $\Pi \in \mathbb{R}^{1 \times n}$, where the i -th column value $\Pi[i] = \rho_i$ is the PageRank value of v_i . The PageRank vector Π satisfies the following equation

$$\Pi = \sum_{t=0}^{\infty} \alpha(1-\alpha)^t \mathbf{s} \cdot \mathbf{P}^t, \quad (3)$$

where $\mathbf{s} \in \mathbb{R}^{1 \times n}$ is the starting vector with the i -th column value $\mathbf{s}[i] = 1/n$ for each $v_i \in \mathcal{V}$. In Eq.(3), $\mathbf{P} \in \mathbb{R}^{n \times n}$ is the probability transition matrix of G where the value in the i -th row and j -th column is $\mathbf{P}[i, j] = 1/d_i$ for $(v_i, v_j) \in \mathcal{E}$ and $\mathbf{P}[i, j] = 0$ otherwise, where d_i is the number of target neighbors of v_i . Correspondingly, a PPR vector $\Pi_i \in \mathbb{R}^{1 \times n}$ w.r.t the user v_i can be derived by setting the starting vector \mathbf{s} of Eq.(3) to an one-hot vector where $\mathbf{s}[i] = 1$ for v_i and $\mathbf{s}[i] = 0$ otherwise. The value in the j -th column $\Pi_i[j] = \pi_{i,j}$ is the PPR value of v_j w.r.t v_i .

To encapsulate PageRank and PPR into social identity features, we bring the idea from [12] which proposed to measure the group centrality by taking the average centrality scores of in-group members. In particular, we define group PageRank (GPR) and group personalized PageRank (GPPR) as follows.

Definition 4.1 (Group PageRank). Given a graph \mathcal{G} , a target user v_t and a candidate group C , the group PageRank of C is defined as

$$\rho_C = \frac{1}{|C|-1} \sum_{v_j \in C \setminus v_t} \rho_j.$$

Definition 4.2 (Group Personalized PageRank). Given a graph \mathcal{G} , a target user v_t and a candidate group C , the group personalized PageRank of C w.r.t v_t is defined as

$$\pi_{t,C} = \frac{1}{|C|-1} \sum_{v_j \in C \setminus v_t} \pi_{t,j}.$$

To understand the intuitions behind GPR and GPPR, the PageRank ρ_j (resp. PPR $\pi_{t,j}$) in Definition 4.1 (resp. Definition 4.2) can be interpreted as measuring the global importance and social status of in-group member v_j [44, 52] (resp. relative importance of v_j w.r.t v_t [54]) by performing the RWR as illustrated in Section 2.2. More concrete, ρ_j represents the probability that an RWR starting from a randomly-selected node stops at v_j . In the meantime, $\pi_{t,j}$ also represents the probability that an RWR stops at v_j , but the starting node is the given node v_t . Thus, GPR and GPPR can indicate the group importance in terms of the probability that a specific RWR terminates in a given group.

Another possible definition for group importance is to replace (personalized) PageRank with the out-degree centrality. However, compared with PageRank, the out-degree centrality d_i of a node v_i only considers the one-hop structural information surrounding v_i , which fails to extensively capture the importance of v_i in the whole graph. Besides the above definitions, [12] also suggests aggregating importance scores of in-group members by summation. Nevertheless, grouping by summation imports the bias from the group scale and yields a misleading high GPR and GPPR scores for the group, in which many users contain but the importance of each is tiny, as justified in Section 5.2.

4.3 Solidarity, Self-stereotyping and In-group Homogeneity

This part proposes user-group tightness (UGT) to seamlessly encapsulate solidarity and self-stereotyping factors, and proposes intra-group tightness (IGT) to quantify the in-group homogeneity factor. More concrete, given a graph \mathcal{G} and a node pair (v_i, v_j) , we use the tie strength $w_{i,j}$ (resp. the similarity $\delta_{i,j}$) as the backbone for the solidarity factor (resp. self-stereotyping and homogeneity factors). W.l.o.g, the similarity score $\delta_{i,j} \in [0, 1]$ is defined as the cosine similarity of the Node2Vec representations [19] of (v_i, v_j) , which is further normalized to the range of $[0, 1]$ by the min-max scaler. In what follows, we formally define UGT and IGT as follows.

Definition 4.3 (User-Group Tightness). Given a graph \mathcal{G} , a target user v_t and a candidate group C , the UGT score of C is defined as

$$\phi_{t,C} = \frac{1}{\sum_{v_j \in C \setminus v_t} w_{t,j}} \sum_{v_j \in C \setminus v_t} w_{t,j} \cdot \delta_{t,j}.$$

Definition 4.4 (Intra-Group Tightness). Given a graph \mathcal{G} a target user v_t and a candidate group C , the IGT score of C is defined as

$$\psi_C = \frac{1}{|C|-1} \sum_{v_j \in C \setminus v_t} \phi_{j,C \setminus v_t}.$$

The intuition behind UGT (resp. IGT) can be interpreted as the averaged attractive spring-like forces between the target user and other group members (resp. among other group members), if the

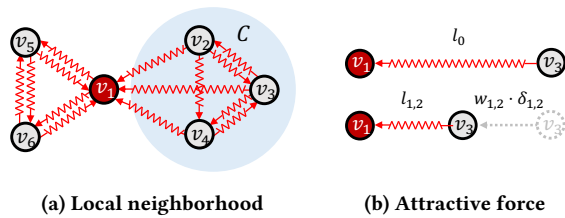


Figure 3: Spring-like local neighborhood and attractive force.

source neighborhood of the target user is abstracted as a force system. This abstraction is widely adopted in graph drawing [14, 16]. Take Figure 3(a) as an example. The source neighborhood of v_1 is represented as a spring network where each edge is assumed as a spring, then the tie strength $w_{i,j}$ can be regarded as the stiffness constant (*i.e.*, physical strength) of the spring on $(v_i, v_j) \in \mathcal{E}$. Let $l_{i,j} \in [0, 1]$ be the natural length (when no force is exerted) of a given spring on (v_i, v_j) , which is equivalent to the theoretical distance between v_i and v_j on the graph, as clarified in [16]. Suppose that all springs are set to initial length 1, then the similarity $\delta_{i,j} = 1 - l_{i,j}$ indicates the displacement of the spring on (v_i, v_j) . According to Hooke’s law, as shown in Figure 3(b), $w_{i,j} \cdot \delta_{i,j}$ is equal to the attractive force exerted by (v_i, v_j) . In contrast, group-level competitors in Section 2.2 only consider the tie strength [2, 8, 43] or the in-group connectivity by the edge density [32, 34], whereas the homogeneity of group members is ignored.

5 EXPERIMENTS

In this part, we first elaborate on the experimental settings, and then evaluate the performance of behavior prediction and target recommendation tasks by employing the proposed measures. All of our experiments are conducted on an in-house cluster consisting of hundreds of machines, each of which runs CentOS, and has 16GB memory and 12 Intel Xeon Processor E5-2670 CPU cores.

5.1 Experimental Settings

Datasets. We use 3 friendship-enhancing event datasets from Tencent’s first person shooter (FPS) and multiuser online battle arena (MOBA) games. A given event, whose procedure is explained in Section 2.3, takes the snapshot of \mathcal{G} before the release time as the input graph, since \mathcal{G} for a particular online game evolves when new users are registered, or friendships are modified. We select events *FPS* and *MOBA-A* as datasets to evaluate the performance on target adoption and source invitation behavior predictions, and events *MOBA-A* and *MOBA-B* to evaluate the performance on target recommendation. The statistics of the graph snapshot, source, and target sets for events are summarized in Table 2. All datasets have been anonymized to avoid any leakage of privacy information.

Competitors. We compare the proposed SIT-based measures with 8 representative prior ones as mentioned in Section 2.2: (i) *individual-level measures*: tie strength (Tie) [13], number of common neighbors (COM) [1], PPR [30], cosine and euclidean similarity between Node2Vec representations (N2V(cos) and N2V(euc)) [19]; (ii) *group-level measures*: structural diversity (#CC) [50], user-group tie strength

Table 2: Dataset statistics ($M = 10^6$, $B = 10^9$).

Dataset	$ \mathcal{V} $	$ \mathcal{E} $	$ \mathcal{S} $	$ \mathcal{T} $
<i>FPS</i>	77.2M	1.1B	33.5M	43.6M
<i>MOBA-A</i>	111.0M	4.5B	111.0M	94.7M
<i>MOBA-B</i>	130.2M	6.5B	120.5M	99.7M

(GT) [2] and group edge density (GD) [32]. All measures are computed based on the network structure as introduced in Table 2. Notice in Tencent’s MOBA and FPS online gaming platforms that the edge weight between a pair of users is described by the *intimacy* score, which records the number of historical activities/interactions from one to the other, *e.g.*, co-playing, gifting, etc. For a fair comparison, we employ intimacy values to measure the tie strength for all related SIT-based measures and competitors. Furthermore, we employ distributed frameworks in [23] and [24, 25] to compute PPR and N2V-based measures, respectively, and set the final embedding dimension of the latent vector to 200. To accomplish downstream tasks, we treat our proposal called SIT (including #CC, GS, GPR, GPPR, UGT, and IGT) or competing measures as input features for XGBoost as mentioned in Section 3, in which the parameters are set following the original paper [7].

5.2 Behavior Prediction

In this section, we separately train two sets of XGBoost models to evaluate the performance of the aforementioned methods in predicting target adoption and source invitation behaviors. Notice that user behaviors can be interfered with by the underlying exposure strategies, which will be illustrated in Section 5.3. To eliminate the bias, we focus on randomly exposed source-target pairs, in which we select the pairs with invitation or adoption behaviors as positive data instances and also randomly select an equal number of pairs without behaviors as negative data instances [55]. Finally, we obtain 116.3 (resp. 12.5) thousands of data instances for invitation (resp. adoption) prediction on *FPS*, and 776.4 (resp. 39.1) thousands of instances for invitation (resp. adoption) prediction on *MOBA-A*, where 80% are used for training and 20% for testing. To evaluate the effectiveness, we repeat each approach 3 times and report the average result under conventional metrics: Area Under Curve (AUC), accuracy and F1 score. In the following experiments, we compare our proposal called SIT with each competitor, every single dimension of SIT and their variants in order, followed by analyzing each SIT dimension. Note that we first show the results about target adoptions and then those about source invitations, as the primary goal of SIT to understand how targets are influenced.

Overall Performance. As illustrated in Table 3, our proposed SIT consistently outperforms other competitors in both *FPS* and *MOBA-A* datasets. Specifically, while predicting target adoption behaviors, SIT is 11.8%, 3.5%, 2.5% (resp. 23.2% 12.6% 6.7%) better than the best competitor in terms of AUC, accuracy, and F1 score in *FPS* (resp. *MOBA-A*). For invitation behaviors, SIT is 16.2%, 12.3%, 16.7% (resp. 12.0%, 8.1%, and 8.7%) better than the best competitor in terms of AUC, accuracy, and F1 score in *FPS* (resp. *MOBA-A*). The above results reflect that involving group information by SIT is necessary for understanding user behaviors in friendship-enhancing events. Regarding competitors, Tie is the best-performing competitor in

Table 3: Comparison of existing measures and proposed SIT measure (the best is colored and the second best is underlined).

Measure	Adoption						Invitation					
	FPS			MOBA-A			FPS			MOBA-A		
	AUC	Accuracy	F1 score	AUC	Accuracy	F1 score	AUC	Accuracy	F1 score	AUC	Accuracy	F1 score
Tie	<u>0.7154</u>	<u>0.6965</u>	0.6554	<u>0.6017</u>	<u>0.6021</u>	0.3958	0.6072	<u>0.5985</u>	0.4607	0.5361	0.5353	0.2200
COM	0.5488	0.5538	0.5615	0.5667	0.5576	<u>0.6219</u>	0.5456	0.5323	0.4674	0.5332	0.5281	<u>0.5565</u>
PPR	0.6565	0.6036	0.5596	0.5562	0.5388	0.4447	<u>0.6289</u>	0.5972	<u>0.5786</u>	<u>0.5846</u>	0.5589	0.5467
N2V(cos)	0.6976	0.6610	<u>0.7171</u>	0.5808	0.5626	0.5420	0.5608	0.5537	0.5683	0.5770	<u>0.5630</u>	0.5426
N2V(euc)	0.7076	0.6652	0.7091	0.5664	0.5566	0.5390	0.5679	0.5588	0.5628	0.5739	0.5585	0.5375
GT	0.6985	0.6572	0.6004	0.6295	0.5959	0.5777	0.5738	0.5652	0.3988	0.5397	0.5297	0.4875
GD	0.6077	0.5736	0.5269	0.6039	0.5728	0.5908	0.5811	0.5507	0.4490	0.5674	0.5508	0.4991
SIT	0.7995	0.7206	0.7350	0.7410	0.6780	0.6638	0.7307	0.6719	0.6754	0.6550	0.6086	0.6047

Table 4: Comparison of different SIT variants (the best is colored and the second best is underlined).

Measure	Adoption						Invitation					
	FPS			MOBA-A			FPS			MOBA-A		
	AUC	Accuracy	F1 score	AUC	Accuracy	F1 score	AUC	Accuracy	F1 score	AUC	Accuracy	F1 score
#CC	0.6153	0.5897	0.5378	0.5452	0.5288	0.4136	0.6091	0.5820	0.5392	0.5790	0.5551	0.5662
GS	0.5866	0.5703	0.5323	0.5860	0.5720	0.6182	0.5669	0.5448	0.4827	0.5655	0.5510	0.5266
GPR	0.6509	0.6071	0.6150	<u>0.6567</u>	<u>0.6149</u>	<u>0.6403</u>	<u>0.6817</u>	<u>0.6314</u>	<u>0.6325</u>	<u>0.6382</u>	<u>0.5978</u>	<u>0.6044</u>
-sum	0.5750	0.5538	0.5702	0.5651	0.5469	0.5798	0.6074	0.5699	0.4978	0.5783	0.5531	0.5545
GPPR	0.6337	0.5990	0.5630	0.5641	0.5405	0.4433	0.6258	0.5922	0.5747	0.5895	0.5603	0.5902
-sum	0.6230	0.5727	0.5749	0.5874	0.5671	0.6217	0.5771	0.5472	0.5634	0.5683	0.5493	0.5545
UGT	<u>0.7427</u>	<u>0.6965</u>	0.7249	0.5634	0.5482	0.5061	0.5969	0.5772	0.5793	0.5684	0.5525	0.5345
-euc	0.7233	0.6857	<u>0.7266</u>	0.5492	0.5345	0.5199	0.5844	0.5643	0.5761	0.5650	0.5505	0.5402
-sum	0.7374	0.6621	0.6145	0.6380	0.6035	0.5793	0.5920	0.5695	0.4477	0.5574	0.5378	0.5695
IGT	0.5622	0.5425	0.4388	0.6296	0.6047	0.5487	0.5372	0.5236	0.3694	0.5642	0.5496	0.4616
-euc	0.5645	0.5650	0.4670	0.5733	0.5569	0.6265	0.5344	0.5239	0.3617	0.5681	0.5521	0.4660
-sum	0.5638	0.5560	0.5008	0.5915	0.5708	0.6058	0.5620	0.5373	0.4605	0.5705	0.5531	0.5170
#CC-GS	0.6698	0.6180	0.6038	0.6101	0.5857	0.6252	0.6306	0.5978	0.5808	0.5972	0.5665	0.5550
SIT	0.7995	0.7206	0.7350	0.7410	0.6780	0.6638	0.7307	0.6719	0.6754	0.6550	0.6086	0.6047

terms of AUC and accuracy scores during the adoption prediction, which indicates that source-target pairs with more historical interactions (*i.e.*, intimacy) are more likely to interact again. Correspondingly, AUC and accuracy scores of GT and GD are also comparable due to the usage of intimacy values. Furthermore, the F1 score of other single-level measures (PPR, COM or N2V(cos)) is the second best on both datasets w.r.t. adoption and invitation predictions. Regarding the structural diversity (#CC), we evaluate its performance in the ablation study, as it is a dimension of SIT.

Ablation study. As illustrated in Table 4, we find that each dimension *i.e.*, #CC, GS, GPR, GPPR, UGT, or IGT, is less effective than SIT. For both behaviors and datasets, UGT or GPR performs as the best competitor in all evaluation metrics, compared with both other SIT-based dimensions and all competitors in Table 3. Regarding IGT, it is less effective than UGT and GPR, but is still comparable to the existing measures in Table 3. For instance, the IGT score is 4.3%, 5.6% better than the group-level competitor GD in terms of AUC and accuracy for adoption behaviors of MOBA-A, as IGT additionally involves the similarities among in-group members.

Notice that both #CC and GS scores are analyzed in the work of structural diversity [50] and can be treated as competitors of other SIT-based measures. In particular, we assemble both measures into XGBoost and denote the model instance as #CC-GS, whose scores can extensively outperform those of each single measure #CC and GS as reported in Table 4. However, #CC-GS are still worse than some single dimensions (*i.e.*, GPR or UGT) and the proposed SIT. Regarding the reasons for the performance of each dimension in SIT, we leave them in the analysis parts.

Factor formulation. As mentioned in Section 4, other possible formulations of proposed measures exist. Thus, except for the averaged aggregation operation and cosine similarity, we also evaluate the performance of variant measures with summation operation and euclidean similarity (denoted as sum and euc, respectively) in Table 4. To derive the euclidean similarity, we first normalize the euclidean distance between each node pair by the min-max scaler as the cosine similarity does, and then take the one minus normalized distance as the corresponding similarity. In particular, employing the summation operation is equivalent to hybrid the

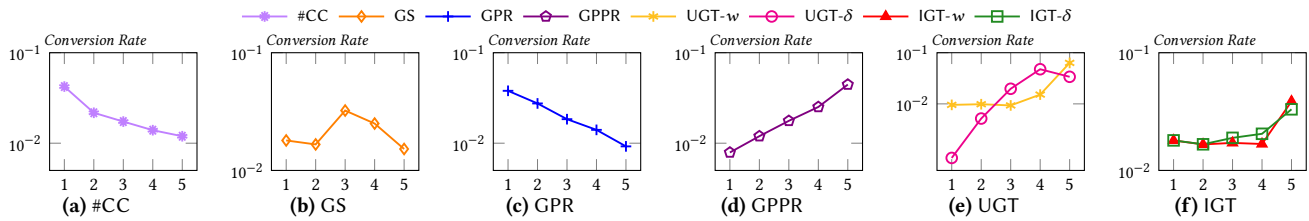


Figure 4: Conversion probability of adoption behaviors conditioned on each SIT-based measure in FPS.

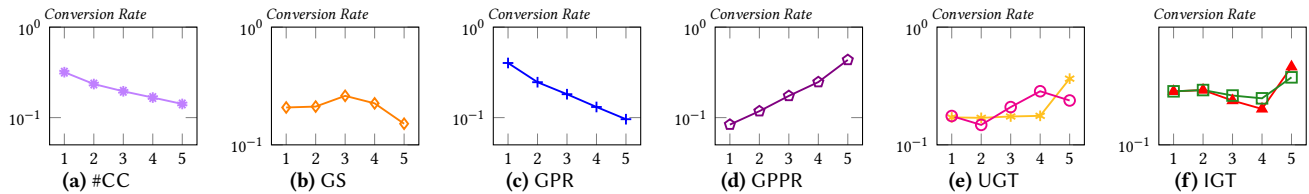


Figure 5: Conversion probability of invitation behaviors conditioned on each SIT-based measure in FPS.

original measure with the factor of GS, which may dampen the original measure (*i.e.*, GPR, GPPR, UGT or IGT) because of the weak correlation between GS and user behaviors, which will be shown in Figures 4-5. For example, the result quality of the summation of GPR is consistently worse than that of GPR on both behaviors and datasets. As for different similarity measures w.r.t node representations, the effectiveness of utilizing cosine and euclidean similarity is comparable. In particular, the F1 scores of euclidean-based UGT and IGT are usually better than those of cosine-based. Meanwhile, negative results are reported regarding AUC and accuracy scores.

Conversion probability conditioned on SIT dimensions. We investigate the relation between user behaviors and the aforementioned SIT-based measures: #CC, GS, GPR, GPPR, UGT, and IGT. Due to space limitations, we only report the analysis results on FPS, and the rest datasets derive similar results. Following the setting of [13], we employ the density-based discretization to convert the scores of SIT-based measures into five different levels, where a higher level indicates a larger score. Moreover, we use *conversion probability* [13, 50] to evaluate how the user acts. In particular, given a level of certain SIT measure, the adoption (*resp.* invitation) conversion probability is the fraction of source-target pairs existing adoption (*resp.* invitation) behaviors over all pairs in this level.

We report the adoption and invitation conversion probabilities w.r.t each measure in Figure 4 and Figure 5, respectively. We can find an evident correlation between each behavior and each proposed measure except for GS. In terms of #CC, contradicting the well-accepted insight of structural diversity [13, 46, 50], Figures 4(a) and 5(a) show that the target user with less number of candidate groups (*i.e.*, a less diverse neighborhood) is more likely to be invited and further adopt this invitation. Recently, this phenomenon also emerged from other user behaviors, especially in Tencent instant messaging platform [34, 55], but the underlying reason is still unclear. In contrast, SIT can explain this as that a greater multi-membership makes the social identification phase of a target challenging due to the potential role conflict [28]. Regarding the effect of group size, we can find in Figures 4(b) and 5(b) that the engagement willingness of users will be facilitated as the group

Table 5: SIT feature importance w.r.t adoption behavior.

	#CC	GS	GPR	GPPR	UGT	IGT
FPS	0.1413	0.1179	0.2291	0.1713	0.2191	0.1212
MOBA-A	0.1561	0.1345	0.1958	0.1733	0.2139	0.1264

Table 6: SIT feature importance w.r.t invitation behavior.

	#CC	GS	GPR	GPPR	UGT	IGT
FPS	0.1404	0.0809	0.2081	0.2221	0.2271	0.1214
MOBA-A	0.1522	0.1002	0.2540	0.1819	0.1907	0.1210

scale increases, but subsequently hurdled if still going up. Notice that the source-target pair inside a medium-size group (*i.e.*, level 3) is more likely to send an invitation and respond.

Recall in Section 4 that GPR and GPPR measure the factors of social standing and centrality from the global and personalized view, respectively. The results in Figures 4(d) and 5(d) show that the target user is more likely to choose a relatively important group, which confirms the claim of the centrality factor in SIT. However, Figures 4(c) and 5(c) illustrate opposing results w.r.t GPR. The reason is that, by definition, a user with a high PageRank value tends to have more neighbors, which consequently means each target friend has less probability of being exposed by the random algorithm.

To evaluate the tie strength and similarity parts in UGT, we separate them as UGT-w and UGT- δ respectively, in which UGT-w measures the averaged intimacy values and UGT- δ measures the averaged cosine similarity between the Node2Vec representations. The same operation is also performed for IGT. As illustrated in Figures 4(e) and 5(e), the behavior conversion probabilities increase as the levels of UGT-w and UGT- δ go up, which match the factors of solidarity and self-stereotyping in SIT. The same tendency for IGT is shown in Figures 4(f) and 5(f). However, the correlation regarding IGT is relatively weak as it fails to provide intra-group information for the group with only one member.

Importance of SIT dimensions. XGBoost supports measuring the importance of each input feature, where the importance indicates the frequency that a feature is leveraged to split the data

Table 7: Online performance in MOBA-A.

Measure	Tie	COM	PPR	N2V(euc)	SIT
E2E rate	0.1018	0.0958	0.1066	0.0739	0.1431

Table 8: Online performance in MOBA-B.

Measure	Tie	PPR	SIT
E2E rate	0.1152	0.1218	0.1384

across all trees. Table 5 and Table 6 report the feature importance of each proposed SIT-based measure w.r.t adoption and invitation behaviors, respectively. In particular, GPR, GPPR, and UGT have the top-3 highest importance w.r.t both behaviors, which indicates that social standing, centrality, solidarity, and self-stereotyping are more critical SIT factors. As for GS and IGT, they perform as the bottom-2 important features. This is because they are less correlated with user behaviors, as illustrated in Figures 4-5. In contrast, #CC is also less important even though it has quite strong correlation to user behaviors. This is due to that #CC is a stand-alone measure for the target user and hence incurs indistinguishable values for source-target pairs with the same target user. To summarize, the aforementioned results of the correlation and importance analysis can explain each SIT dimension’s performance in Table 4.

5.3 Target Recommendation

Deployment setups. Recall in Section 2.3 that a source user can only select the target users from a feed window of limited size k during the second step of a friendship-enhancing event. Therefore, judiciously exposing a subset of target friends for the source v_i is pivotal to the event’s performance. Motivated by this, the target recommendation task (Problem 2) is to find k target users w.r.t the source to boost the overall engagement of sources and targets. Here, we inherit the XGBoost model, which takes the aforementioned individual measures (*i.e.*, Tie (measured by intimacy), COM, PPR, and N2V(euc)) or SIT measures as the input and is trained based on the historical events. We sort the predicted value $\hat{y}_{i,j}$ w.r.t each source v_i in a descending order and select the top- k target nodes to recommend. To evaluate the proposed SIT measures, we conduct the online A/B testing that randomly assigns a fraction of live traffic to XGBoost models with individual-level measures as treatment groups. Initially, each measure is computed based on the graph instance ahead of the event (see Table 2). Afterward, each measure is updated daily by using the latest graph snapshot.

Overall performance. We evaluate the effectiveness of different online random trials by the end-to-end (E2E) rate, which considers the overall engagement of sources and targets. In particular, E2E rate equals the fraction of target friends adopting the invitations over the source users seeing the event. The higher E2E rate indicates better quality. As illustrated in Table 7, the proposed solution SIT has the highest E2E rate on MOBA-A. Specifically, SIT is 34.2% better than the best-performed treatment approach PPR. Furthermore, we can find that PPR and Tie can outperform the other two competitors. Motivated by this, we conduct another online trial on MOBA-B by comparing SIT with the best two measures PPR and Tie. Specifically,

Table 8 reports that SIT is 13.6% better than the best-performed treatment approach PPR.

6 ADDITIONAL RELATED WORKS

This part briefly reviews other TFC measures and the related sociological theory. In particular, prior TFC measures can be explained by either *selection* or *influence* mechanism in the homophily principle [10]. The selection indicates that people tend to form friendships with others with similar characteristics. In contrast, the influence can be treated as the inverse of selection, claiming that people may modify characteristics to conform to their friends. We categorize TFC measures based on the taxonomy of selection and influence.

Regarding the selection, conventional TFC measures utilize the variant based on common neighborhood, *e.g.*, Adamic/Adar score [1]. Furthermore, the selection can be measured by the similarity between node embeddings, *e.g.*, [19, 31, 36, 48, 51, 54], which essentially rely on the factorization of a matrix of the TFC measure as mentioned in Section 2.2 [33]. Regarding social influence, the related measures mainly focus on the local structure of a user. Burt [4] proposes to model social influence from cohesion and structural equivalence perspectives. In particular, cohesion describes the direct influence measured by tie strength and the number of influenced friends [13, 18]. The structural equivalence describes the indirect influence between users and is measured by tie embeddedness [10]. Recently, many learning-based methods [17, 22, 37] are proposed to predict the influence probability based on historical cascades, among which [22] leverages another psychological concept called conformity to measure the social influence.

In contrast, SIT is a group-level psychological framework, in which each dimension can reflect either selection or influence in the homophily principle. For instance, the factors of self-stereotyping and in-group homogeneity are coherent with the selection mechanism. Therefore, SIT can also provide theoretical support for the aforementioned measures and their integration, *e.g.*, [37].

7 CONCLUSIONS

In this work, we propose measures for friendship closeness based on the social identity theory, a fundamental concept in social psychology, and conduct extensive experiments and analysis on 3 Tencent’s online gaming datasets. Compared with 8 state-of-the-art methods, our proposal achieves the highest effectiveness on both behavior prediction and target recommendation tasks. Regarding future works, we will employ presented measures for other scenarios, *e.g.*, modeling the cascade of social influence and suggesting strangers to enrich users’ friendships in the who-to-follow service.

ACKNOWLEDGMENTS

This work was supported by Proxima Beta (Grant No. A-8000177-00-00) and the Guangdong Provincial Key Laboratory (Grant No. 2020B121201001). Bo Tang is also affiliated with the Research Institute of Trustworthy Autonomous Systems, Shenzhen, China and Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, China.

REFERENCES

- [1] Lada A Adamic and Eytan Adar. 2003. Friends and neighbors on the web. *Social networks* 25, 3 (2003), 211–230.
- [2] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2014. Who to follow and why: link prediction with explanations. In *SIGKDD*. 1266–1275.
- [3] Marilyn B Brewer and Wendi Gardner. 1996. Who is this “We”? Levels of collective identity and self representations. *Journal of personality and social psychology* 71, 1 (1996), 83.
- [4] Ronald S Burt. 1987. Social contagion and innovation: Cohesion versus structural equivalence. *American journal of Sociology* 92, 6 (1987), 1287–1335.
- [5] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *TKDE* 30, 9 (2018), 1616–1637.
- [6] James E Cameron. 2004. A three-factor model of social identity. *Self and identity* 3, 3 (2004), 239–262.
- [7] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *SIGKDD*. 785–794.
- [8] Abir De, Sourangshu Bhattacharya, Sourav Sarkar, Niloy Ganguly, and Soumen Chakrabarti. 2016. Discriminative link prediction using local, community, and global signals. *TKDE* 28, 8 (2016), 2057–2070.
- [9] Bertjan Doosje, Nyla R Branscombe, Russell Spears, and Antony SR Manstead. 1998. Guilty by association: When one’s group has a negative history. *Journal of personality and social psychology* 75, 4 (1998), 872.
- [10] David Easley, Jon Kleinberg, et al. 2010. *Networks, crowds, and markets*. Vol. 8. Cambridge university press Cambridge.
- [11] Alessandro Epasto, Silvio Lattanzi, Vahab Mirrokni, Ismail Oner Sebe, Ahmed Taei, and Sunita Verma. 2015. Ego-net community mining applied to friend suggestion. *PVLDB* 9, 4 (2015), 324–335.
- [12] Martin G Everett and Stephen P Borgatti. 1999. The centrality of groups and classes. *The Journal of mathematical sociology* 23, 3 (1999), 181–201.
- [13] Zhanpeng Fang, Xinyu Zhou, Jie Tang, Wei Shao, Alvis Cheuk M Fong, Longjun Sun, Ying Ding, Ling Zhou, and Jarder Luo. 2014. Modeling paying behavior in game social networks. In *CIKM*. 411–420.
- [14] Thomas MJ Fruchterman and Edward M Reingold. 1991. Graph drawing by force-directed placement. *SP&E* 21, 11 (1991), 1129–1164.
- [15] Samuel L Gaertner and John F Dovidio. 2014. *Reducing intergroup bias: The common ingroup identity model*. Psychology Press.
- [16] Emden R Gansner, Yehuda Koren, and Stephen North. 2004. Graph drawing by stress majorization. In *GD*. 239–250.
- [17] Manuel Gomez-Rodriguez, David Balduzzi, and Bernhard Schölkopf. 2011. Uncovering the temporal dynamics of diffusion networks. In *ICML*.
- [18] Mark Granovetter. 1978. Threshold models of collective behavior. *American journal of sociology* (1978).
- [19] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *SIGKDD*. 855–864.
- [20] Glen Jeh and Jennifer Widom. 2002. Simrank: a measure of structural-context similarity. In *SIGKDD*. 538–543.
- [21] Colin Wayne Leach, Martijn Van Zomeren, Sven Zebel, Michael LW Vliek, Sjoerd F Pennekamp, Bertjan Doosje, Jaap W Ouwerkerk, and Russell Spears. 2008. Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification. *Journal of personality and social psychology* 95, 1 (2008), 144.
- [22] Hui Li, Hui Li, and Sourav S Bhowmick. 2020. Chassis: Conformity meets online information diffusion. In *SIGMOD*. 1829–1840.
- [23] Wenqing Lin. 2019. Distributed algorithms for fully personalized pagerank on large graphs. In *TheWebConf*. 1084–1094.
- [24] Wenqing Lin. 2021. Large-Scale Network Embedding in Apache Spark. In *SIGKDD*. 3271–3279.
- [25] Wenqing Lin, Feng He, Faqiang Zhang, Xu Cheng, and Hongyun Cai. 2020. Initialization for Network Embedding: A Graph Partition Approach. In *WSDM*. 367–374.
- [26] Siqiang Luo, Xiaokui Xiao, Wenqing Lin, and Ben Kao. 2019. Efficient batch one-hop personalized pageranks. In *ICDE*. 1562–1565.
- [27] Siqiang Luo, Xiaokui Xiao, Wenqing Lin, and Ben Kao. 2020. BATON: Batch One-Hop Personalized PageRanks with Efficiency and Accuracy. *IEEE Trans. Knowl. Data Eng.* 32, 10 (2020), 1897–1908.
- [28] Stephen R Marks. 1977. Multiple roles and role strain: Some notes on human energy, time and commitment. *American sociological review* (1977), 921–936.
- [29] Penelope J Oakes, S Alexander Haslam, and John C Turner. 1994. *Stereotyping and social reality*. Blackwell Publishing.
- [30] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web*. Technical Report. Stanford InfoLab.
- [31] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *SIGKDD*. 701–710.
- [32] Hemant Purohit, Yiye Ruan, David Fuhry, Srinivasan Parthasarathy, and Amit Sheth. 2014. On understanding the divergence of online social group discussion. In *ICWSM*, Vol. 8. 396–405.
- [33] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. 2018. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *WSDM*. 459–467.
- [34] Jiezhong Qiu, Yixuan Li, Jie Tang, Zheng Lu, Hao Ye, Bo Chen, Qiang Yang, and John E Hopcroft. 2016. The lifecycle and cascade of wechat social messaging groups. In *TheWebConf*. 311–320.
- [35] Tim Rees, S Alexander Haslam, Pete Coffee, and David Lavallee. 2015. A social identity approach to sport psychology: Principles, practice, and prospects. *Sports medicine* 45, 8 (2015), 1083–1096.
- [36] Sam T Roweis and Lawrence K Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 5500 (2000), 2323–2326.
- [37] Aravind Sankar, Xinyang Zhang, Adit Krishnan, and Jiawei Han. 2020. Inf-vae: A variational autoencoder framework to integrate homophily and influence in diffusion prediction. In *WSDM*. 510–518.
- [38] Daan Scheepers and Naomi Ellemers. 2019. Social identity theory. In *Social psychology in action*. 129–143.
- [39] Joseph Seering, Felicia Ng, Zheng Yao, and Geoff Kaufman. 2018. Applications of social identity theory to research and design in computer-supported cooperative work. *PCHI* 2, CSCW (2018), 1–34.
- [40] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD* 19, 1 (2017), 22–36.
- [41] Sam D Sieber. 1974. Toward a theory of role accumulation. *American sociological review* (1974), 567–578.
- [42] Bernd Simon. 1992. The perception of ingroup and outgroup homogeneity: Reintroducing the intergroup context. *European review of social psychology* 3, 1 (1992), 1–30.
- [43] Junho Song, Hyekyoung Park, Kyungsik Han, and Sang-Wook Kim. 2020. Do You Really Like Her Post?: Network-Based Analysis for Understanding Like Activities in SNS. In *CIKM*. 2221–2224.
- [44] Xiaodan Song, Yun Chi, Koji Hino, and Belle Tseng. 2007. Identifying opinion leaders in the blogosphere. In *CIKM*. 971–974.
- [45] Sucheta Soundarajan and John Hopcroft. 2012. Using community information to improve the precision of link prediction methods. In *WWW*. 607–608.
- [46] Jessica Su, Krishna Kamath, Aneesh Sharma, Johan Ugander, and Sharad Goel. 2020. An experimental study of structural diversity in social networks. In *ICWSM*, Vol. 14. 661–670.
- [47] Henri Tajfel and John C Turner. 2004. The social identity theory of intergroup behavior. In *Political psychology*. 276–293.
- [48] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *TheWebConf*. 1067–1077.
- [49] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. 2006. Fast random walk with restart and its applications. In *ICDM*. 613–622.
- [50] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. 2012. Structural diversity in social contagion. *PNAS* 109, 16 (2012), 5962–5966.
- [51] Daixin Wang, Peng Cui, and Wenwu Zhu. 2016. Structural deep network embedding. In *SIGKDD*. 1225–1234.
- [52] Liang Wu, Xia Hu, and Huan Liu. 2016. Relational learning with social status analysis. In *WSDM*. 513–522.
- [53] Jaewon Yang and Jure Leskovec. 2015. Defining and evaluating network communities based on ground-truth. *KIS* 42, 1 (2015), 181–213.
- [54] Renchi Yang, Jieming Shi, Xiaokui Xiao, Yin Yang, and Sourav S Bhowmick. 2020. Homogeneous Network Embedding for Massive Graphs via Reweighted Personalized PageRank. *PVLDB* 13, 5 (2020).
- [55] Fanjin Zhang, Jie Tang, Xueyi Liu, Zhenyu Hou, Yuxiao Dong, Jing Zhang, Xiao Liu, Ruobing Xie, Kai Zhuang, Xu Zhang, et al. 2021. Understanding WeChat User Preferences and “Wow” Diffusion. *TKDE* 1 (2021), 1–1.
- [56] Yuanxing Zhang, Zhuqi Li, Chengliang Gao, Kaigui Bian, Lingyang Song, Shaoling Dong, and Xiaoming Li. 2018. Mobile social big data: Wechat moments dataset, network applications, and opportunities. *IEEE network* 32, 3 (2018), 146–153.