DSCAN for Geo-Social Team Formation

Maryam MahdavyRad, Kalyani Selvarajah, and Ziad Kobti

School of Computer Science, University of Windsor, Windsor, ON, Canada
{mahdavy,kalyanis,kobti}@uwindsor.ca

Abstract. Nowadays, geo-based social group activities have become popular because of the availability of geo-location information. In this paper, we propose a novel Geo-Social Team Formation framework using DSCAN, named DSCAN-GSTF, for impromptu activities, aim to find a group of individuals closest to a location where service requires quickly. The group should be socially cohesive for better collaboration and spatially close to minimize the preparation time. To imitate the real-world scenario, the DSCAN-GSTF framework considers various criteria which can provide effective Geo-Social groups, including a required list of skills, the minimum number of each skill, contribution capacity, and the weight of the user's skills. The existing geo-social models ignore the expertise level of individuals and fail to process a large geo-social network efficiently, which is highly important for an urgent service request. In addition to considering expertise level in our model, we also utilize the DSCAN method to create clusters in parallel machines, which makes the searching process very fast in large networks. Also, we propose a polynomial parametric network flow algorithm to check the skills criteria, which boosts the searching speed of our model. Finally, extensive experiments were conducted on real datasets to determine a competitive solution compared to other existing state-of-the-art methods.

Keywords: Geo-Social Networks · Geo-Social Groups · DSCAN

1 Introduction

Nowadays, geo-based social group activities have become popular because of the availability of geo-location information. In this paper, we propose a novel Geo-Social Team Formation framework using DSCAN, named DSCAN-GSTF, for impromptu activities, aim to find a group of Geo-Social Networks (GeoSNs) is online social networks that allow geo-located information to be shared in real-time. The availability of location acquisition technologies such as GPS and WiFi enables people to easily share their position and preferences to existing online social networks. Here, the preferences can be common interests, behavior, social relationships, and activities. This information is usually derived from a history of an individual's locations and Geo-tagged data, such as location-tagged photos and the place of the current event [27]. Thus, we have several popular GeoSNs such as Facebook, Twitter, Flickr, Foursquare, Yelp, Meetup, Gowalla, and Loopt. Consequently, GeoSNs have drawn significant attention in recent years by researchers on many applications, including finding friends in the vicinity [23, 13], group-based activity planning [4], and marketing [6].

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_44

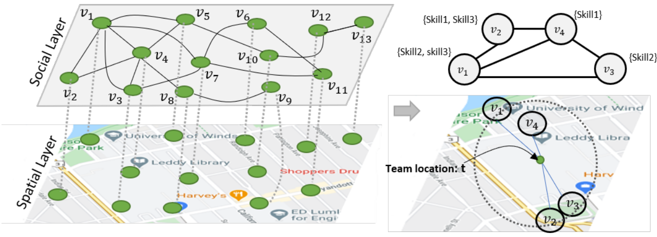


Fig. 1. Identifying individuals for impromptu social activity from a GeoSNs.

An impromptu activity planning is one of the popular motivating applications in GeoSNs search. For example, the COVID19 outbreak is affecting every part of human lives. At the initial stage of COVID19, essential services such as health, finance, food, and safety suffered a lot due to unexpected lockdown because they did not have the required human resources as expected. At the same time, fulfilling societies' requirements are also equally important. Therefore, bringing skilled people to the location where the services with diverse demands were crucial and had become a very challenging process. The location of the individual should be close to the place where service is required, and the individual with the required skills needs to be suitable for services. In this example, a service might require several skills. Additionally, each individual may contribute to as many skills as possible in various activities and might have a specific capacity to be involved in multiple activities.

To support this situation, we highly believe that forming groups from GeoSN is an effective solution. In general, this is called the Geo-social group search problem [18, 19], which aim to identify groups of individual who are socially cohesive and spatially closest to a location [4]. In other words, the Geo-social group should satisfy that the participants are socially close within the group to confirm good communication between each member and spatially closest to the location of the service to bring them as soon as possible. Figure 1 represents a general Geo-Social Network, where the social layer is to show the social connections between individuals, and the spatial layer is to show the locations of the individuals.

Many existing Geo-social group models considered social cohesiveness and spatial closeness to find successful groups. In addition to these two requirements, recently, Chen et al. [4] incorporated a few essential parameters such as the collective capabilities and capacity of each member. However, to the best of our knowledge, none of the existing Geo-social group models considered the weights of the user's skills which assist in choosing the exact qualified individual. Moreover, efficiently searching required keywords that have high expertise, capacity constraints, social constraint, and spatial closeness altogether have not been explored in the existing studies. Forming a search framework that can quickly narrow the search space while preserving the correct result is an NP-Hard. It is an open problem and essential to solve in polynomial time [4].

This paper proposes a novel framework to search efficiently on large GeoSNs while preserving the correct result. First, handling large networks is a time-consuming process. So we adopt a recently proposed methodology, Distributed

Structural Clustering Algorithm for Networks (DSCAN) [20] algorithm to efficiently manage large networks, which is an extension of SCAN [24]. The basic idea of SCAN is to discover clusters, hubs, and outliers included in a given graph. Initially, in our model, all the nodes of a given graph are randomly divided into equal size sub-graphs and distributed into different machines so that the remaining processing can be conducted simultaneously. Then, by employing the skewness-aware edge-pruning method on the sub-graphs, DSCAN eliminates unwanted edges and moves missing neighbors of nodes from one sub-graph to another. Second, producing socially cohesive groups from these sub-graphs is another essential process. So, DSCAN collects the Core nodes with higher structural similarity and creates a set of clusters from these sub-graphs. Parallelly, the set of skills is collected from each cluster and stored on a map. The third requirement is to choose a spatially close group to the location (∇) where the service is required. We pick a node randomly from each cluster and evaluate the geographical distance from ∇. The clusters are ranked based on the distance in ascending order. Then a cluster with the lowest rank is selected and tested to see whether it satisfies the requirements of the query or not. If it does not satisfy, move to the following cluster and test the requirement. This process will be repeated till we find the right cluster. Finally, we propose a new polynomial algorithm based on the parametric flow network [8], which checks the skills requirements of the query and contribution capacity of each individual in the selected cluster while considering the user's skills weights.

Our Contribution: The followings are the summary of our contributions:

1. We propose a Geo-Social Team Formation (GSTF) model by considering the group's collective capabilities as required for the activities while considering the capacity of contribution from each member and expertise level.
2. We utilize the benefits of Distributed Structural Graph Clustering (DSCAN) to manage the large GeoSNs efficiently.
3. We propose a new polynomial searching algorithm based on the parametric flow network, which satisfies Minimum keywords, expertise level, and capacity constraints.

The rest of the paper is organized as follows. Section 2 discusses related existing work. Section 3 defines the problem definitions of our proposed model. Our framework is presented in Section 4. Following that, Section 5 illustrates the experimental setup and the corresponding results. Finally, Section 6 concludes the research idea of this paper with directions for future work.

2 Related Work

Forming a group of individuals for various purposes has been tackled in many different ways. The team formation problem in Social networks was first introduced by Lappas et al.[12]. Later many studies [10, 16] have been conducted by incorporating various parameters which influence the successful formation of teams in several applications, including academic collaborations, healthcare [17], and human resource management. However, many of these studies focused

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_44

highly on minimizing or maximizing some social constraints such as communication cost between members in a team based on their past relationship, profit, and productivity cost.

The concept of GeoSNs services was first introduced by Huang et al. [9]. Many studies have focused on querying geo-social data in order to derive valuable information from both the users' social interactions and current locations [1, 21]. Among these, forming Geo-social groups has taken considerable attention of researchers recently since this aims to identify a set of most suitable individuals for various activities which can be planned or unplanned. The unplanned activities such as groups for various purposes during unexpected events, for example, Wildfire and flooding, are relatively complex than the planned activities such as a group for a party or a game. Much existing research proposed various models for both situations while satisfying social constraints and optimizing spatial proximity [4, 21]. Many of these focused on forming a group that satisfies a single social constraint while optimizing the spatial proximity. But for impromptu activities, in reality, we require individuals who have diverse demands of skills for multiple tasks or services to serve in a specific location. Recently, Chen et al. [4] introduced a novel framework to discover a set of groups that is socially cohesive while spatially closest to a location for diverse demands. Here, the groups of individuals do not necessarily know each other in the past. However, When there is a tie between two members, their model gives higher priority to the individual who is highly cohesive to the team. The concept of multiple social constraints for various activities has already been studied in [17, 15]. However, they considered the frameworks on social networks with known individuals.

Searching cohesive subgroups from a large network is another challenging process in the team formation problem. Structural Clustering Algorithm for Networks (SCAN) algorithm was proposed to detect cohesive subgroups from a network [24]. However, SCAN is a computationally expensive method for a large network because it requires iterative calculation for all nodes and edges. Later, to overcome the limitation with SCAN, many clustering methods have been proposed, such as PSCAN [26], and DSCAN [20]. Since DSCAN is efficient, scalable, and exact, we employ this methodology in our model. To the best of our knowledge, we, for the first time, applied DSCAN in the Geo-Social group search problem.

3 Problem Definition

Given an undirected graph $G = (V, E)$, where V is a set of vertices and E is a set of edges. The Graph G incorporates network structures, spatial information, and textual information. In real networks, vertices are users or people, and edges between them may be friendship or previous collaboration. Additionally, each vertex $v \in V$ includes location information, which can be represented as $\nabla = (v.x, v.y)$, where $v.x$ is latitude and $v.y$ is longitude, and a set of keyword attributes which can be represented as $v.A$. The textual information can be a set of skills $S = \{s_1, s_2, \ldots, s_k\}$ of a vertex $v \in V$, where k is the number of skills that a person is expert in. Along with the skills, a vertex has a set weight $W = \{w_1, w_2, \ldots w_k\}$ to represent how much a person expert in each skill.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_44

Definition 1. ***Query** (Q):* *The query defines the requirements of skills and number of people in each skills. This includes a Geo-location ∇ (latitude (x) and longitude (y)), a set of required skills $S = \{s_1, s_2, \ldots, s_r\}$, a set of required expertness in each skills $P = \{p_1, p_2, \ldots, p_r\}$ and a contribution capacity of an expert c for every query keyword needs to be assigned.*

Definition 2. ***Geo-Social Team** (B):* *For a given location ∇, a set of the required number of experts who satisfies social cohesiveness and spatial closeness is selected from a Geo-Social network G while considering contribution capacity c and person's skill weight in each skill.*

In our model, we exploit the DSCAN to handle larger data efficiently. To understand the concept of DSCAN, the following definition are necessary.

Definition 3. ***Structural Neighborhood** (N_v):* *The structural neighborhood N_v of vertex v can be defined as,*

$$N_v = \{w \in V | (v, w) \in E\} \cup \{v\} \tag{1}$$

Definition 4. ***Structural similarity:*** *The structural similarity $\sigma(v, w)$ between v and w can be defined as,*

$$\sigma(v, w) = |N_v \cap N_w| / \sqrt{|N_v||N_w|} \tag{2}$$

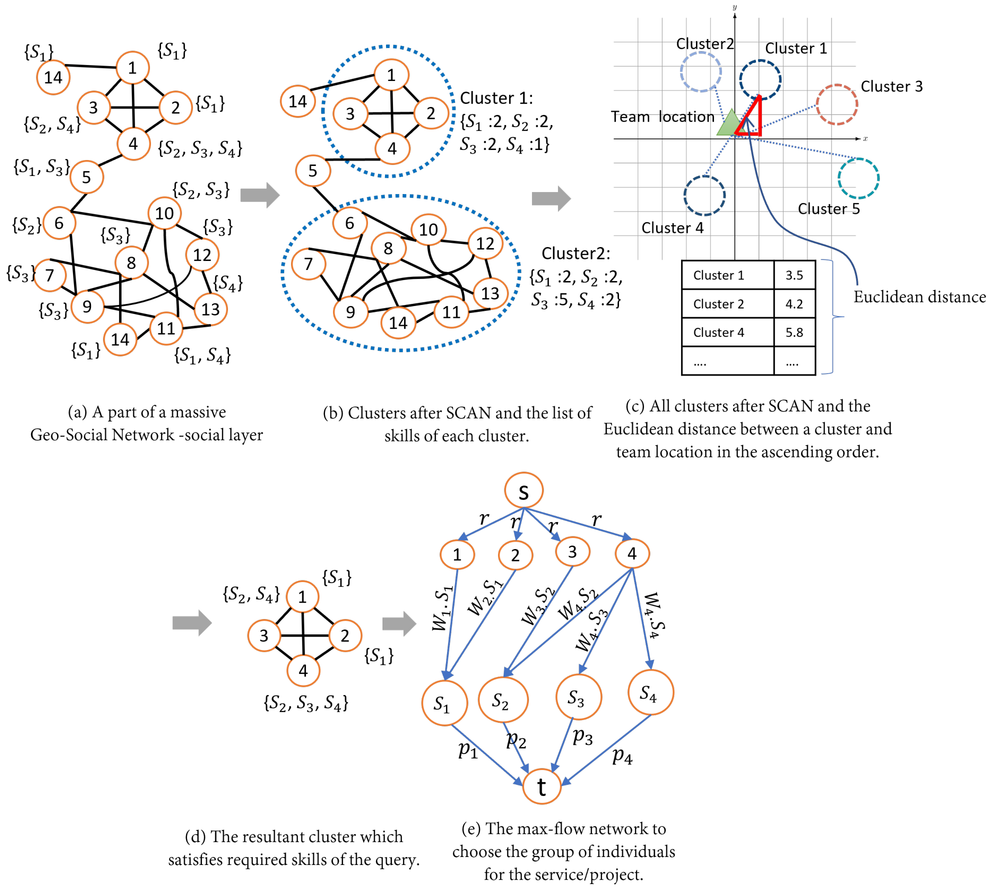
If $\sigma(v, w) \geq \epsilon$, vertex v shares similarity with w and $\epsilon \in \mathbb{R}$ is a density threshold which we assigned.

When a vertex has enough structurally identical neighbors, it becomes the seed of a cluster, named core node. Core nodes have at least μ number of neighbors with a structural similarity $(\sigma(v, w))$ that exceeds the threshold ϵ.

Definition 5. ***Core:*** *For a given ϵ and μ, A vertex $w \in V$ is called a core, iff $N_{w,\epsilon} \geq \mu$. where $N_{w,\epsilon}$ is the set of neighbor nodes of core node w, and structural similarity of $N_{w,\epsilon}$ is greater than ϵ.*

Definition 6. ***Cluster** ($\mathcal{C}_w$):* *Assume node w be a core node. SCAN collects all nodes in $N_{w,\epsilon}$ into the same cluster ($\mathcal{C}_w$) of node w, initially $\mathcal{C}_w = \{w\}$. SCAN outputs a cluster $\mathcal{C}_w = \{v \in N_{u,\epsilon} | u \in \mathcal{C}_w\}$.*

DSCAN Algorithm: When DSCAN [20] receives a graph, it first deploys disjointed subgraphs of the given graph G to distributed memories on multiple machines $M = \{M_1, M_2, \ldots, M_n\}$ for a given a density threshold $\epsilon \in \mathbb{R}$ and a minimum size of a cluster $\mu \in \mathbb{N}$, where n is number of machines. Initially DSCAN randomly moves a set of vertices V_i in subgraph $G_i = (V_i, E_i)$ for each machine M_i. The subgraphs are then processed in a parallel and distributed fashion. Additionally, DSCAN uses edge-pruning based on skewness to improve efficiency further.
Skewness-Aware Edge-Pruning: DSCAN applies $\omega-$skewness edge-pruning to remove unwanted edges and move missing neighbors of nodes from one subgraph to another. Given graph $G = (V, E)$, consider an edge (u, v) be in E and



Cluster 1	3.5
Cluster 2	4.2
Cluster 4	5.8
....	....

(a) A part of a massive Geo-Social Network -social layer

(b) Clusters after SCAN and the list of skills of each cluster.

(c) All clusters after SCAN and the Euclidean distance between a cluster and team location in the ascending order.

(d) The resultant cluster which satisfies required skills of the query.

(e) The max-flow network to choose the group of individuals for the service/project.

Fig. 2. The DSCAN-GSTF Framework: (a) A part of a massive Geo-Social Network-Social layer, (b) The clusters and the list of skills that each cluster satisfies. (c) The ordered distance between a cluster and the team location (d) Selected cluster which satisfies spatial constraint and skill constraint. (e) The max-flow network to choose the successful individuals for the service.

the structural neighborhood of node u is $N_u = \{v \in V | ((u, v) \in E)\} \cup \{u\}$. And $\omega-$skewness for each edge can be defined as,

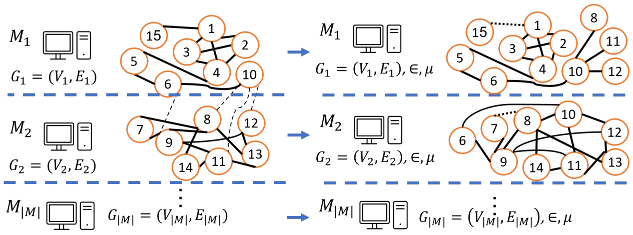
$$\omega(u, v) = min\Big\{\frac{d_u}{d_v}, \frac{d_v}{d_u}\Big\} \tag{3}$$

where $d_i = |N_i|$. If $\omega(u, v) < \epsilon^2$, then the edge (u, v) is considered dissimilar and prune from the graph [20].

4 Our Framework

The Geo-Social team formation framework, DSCAN-GSTF consists of three primary processes.1) Distribute G into multiple machines and perform local clustering. 2) Choose the cluster proximate to the location.3) Apply parametric flow algorithm to select a competent Geo-Social team of experts. We describe each step one by one in the following sections.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_44


M_1
$G_1 = (V_1, E_1)$
M_2
$G_2 = (V_2, E_2)$
$M_{|M|}$
$G_{|M|} = (V_{|M|}, E_{|M|})$
$G_1 = (V_1, E_1), \in, \mu$
$G_2 = (V_2, E_2), \in, \mu$
$G_{|M|} = (V_{|M|}, E_{|M|}), \in, \mu$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

(a) All vertices are randomly divided equally into different machines

(b) The resultant sub-graphs after applying ω-skewness edge-pruning

Fig. 3. The overview of parallel processing by using DSCAN to replace the process of Figure 2 (b).

4.1 Network Distribution

The GeoSNs are very large networks with millions of edges and vertices. We replicated DSCAN [20] framework in our application. A given large network G is randomly divided into equal size of sub-graphs $\{G_1, G_2, \ldots, G_n\}$. We then deploy each sub-graph into separate machine M_i as shown in Figure 3 (a). However, sub-graphs G_i and G_j might have neighbor nodes with higher structural similarity ($\geq \mu$). Those nodes should communicate across machines M_i and M_j. So, DSCAN employs skewness-aware edge-pruning to keep a low communication cost for billion-edge graphs [20], which is shown in Figure 3 (b). The skewness-aware edge-pruning drops unnecessary edges to avoid the unwanted communication cost among the machines and moves missing neighbors of nodes which have a high structural similarity, from one sub-graph to another.

The sub-graphs which are placed in each machine, are again clustered based on the structural similarity [26]. Here, DSCAN finds all core nodes in each sub-graph and constructs clusters with the nodes which have high structural similarities. Additionally, we store the list of skills of each cluster. The example of resultant clusters and the list of skills are displayed in Figure 2 (b).

4.2 Suitable Cluster Selection

From the list of clusters, we then select the clusters that satisfy the skill constraints of the given query. To ensure the spatial proximity, we evaluate the Euclidean distance between each cluster and the location ∇ where the service is required (Figure 2 (c)). We order these distances in ascending order and choose the nearest one that is satisfy the required list of skills as shown in Figure 2 (d). The selected one is then sent to searching algorithms to find a competent geo-social team. We discuss this process in the following section.

4.3 Geo-Social Team Formation

We propose a polynomial searching algorithm based on the parametric flow network [8] to find a competent Geo-Social team (B). We describe the preliminaries of the parametric flow network one by one.

DOI: 10.1007/978-3-031-08757-8_44

Flow Network: A flow network $N_G = (N_V, N_E)$ is a directed graph that contains a source node s, an target node t, a set of middle nodes N_V, and directed edge set N_E. Additionally, each edge has a weight and receive a flow. An edge's weight cannot be exceeded by the amount of flow that passes through it.
Max Flow and min $s - t$ cut: Let's say f is a flow of N_G, the flow across an $s - t$ cut (S, T) divide its nodes into S and T parts so that the sum of the capacities across S and T is minimized. So, the maximum amount of flow moving from an $s - t$ cut in N_G, say $f(S, T)$ is equal to the total weight of the edges in a minimum cut, $\sum_{u \in S, v \in T} f(u, v)$.
Preflow: A PreFlow f on N_G is a real-valued function that satisfies the capacity and anti-symmetric constraints on node pairs. The relaxation of the conservation constraint can be defined as, $\sum_{u \in V(D)} f(u, v) \geq 0, \forall u \in V \backslash \{s\}$
Valid Labelling: A valid labelling h for a preflow f is a function which is attached to the vertices and has positive integers, such that $h(t) = 0$, $h(s) = |N_V(N_G)|$, where $|N_V(N_G)|$ is the number of vertices in network N_G [3]. For every directed edge from node v to u, the relabeling of $h(v) \leq h(u) + 1$ should be created to have a valid flow. In other words, for any node v is a valid labelling if $h(v) \leq min\{h_f(v, t), h_f(v, s) + |V(D)|\}$. The purpose of such labelling $h(v)$ is to estimate the shortest distance from the vertex v to s or t [7].
Calculation of min s-t cut: After running the max-flow algorithm, a minimum cut can be found as follows. For each node $v \in V$, replace $h(v)$ by $min\{h_f(s, v), h_f(t, v) + |N_V(N_G)|\}$. Now the $s - t$ cut is equal to $S = \{v | h(v) \geq |N_V(N_G)|\}$ where the sink partition T is of the minimum size.
Parametric Network Flow: The maximum or minimum value of the flow is determined using a max-flow algorithm based on some criteria. In a parametric-flow network N_R, the capacities on arcs out of s and into t are functions of a real-valued parameter λ, and edges possess the following characteristics [8]. For all $v \neq t$ the cost of the edges from source node to v nodes $C_{(s,v)}(\lambda)$ is a non-decreasing function of λ. Also, for all $v \neq s$ the cost of the edges from v nodes to target node t, $C_{(v,t)}(\lambda)$ is a non-increasing function of λ. And finally, for all $v \neq s$ and $v \neq t$ the cost of the edges from node v to node u, and $C_{(u,v)}(\lambda)$ is a constant. Parametric networks measure maximum flow or minimum cut based on a particular parameter value λ.
Triangle in graphs: A triangle in G is a cycle of length 3. A triangle generated on vertices $u, v, w \in V(G)$ is denoted as $Tri(uvw)$.
Context weighted density (CW)**:** In the selected subgraph $H \subset G$ which satisfies the requirement of query Q, vertices that are related to the query Q may be loosely or densely connected. To balance both these situation, we decided to evaluate context weighted density, (CW) so that we can have cohesive group [22]. The context weighted density, (CW) can be calculated with the use of both weighted triangle density and weighted edge density. For a given edge $(u, v) \in E(H)$, $(u, v, w) \in Tri(H)$, the context scores can be defined as below,

$$\text{Edge context score:} w(e(u, v)) = |Q \cap A(u)| + |Q \cap A(v)| \tag{4}$$

$$\text{Triangle context score:} W(T(u, v, w)) = \sum_{e \in \{(u,v),(u,w),(v,w)\}} w(e) \tag{5}$$

DOI: 10.1007/978-3-031-08757-8_44

Algorithm 1 Skills Query Search Algorithm

```
Input: cluster $H \in G$, Query $Q$

1: $H_0 \leftarrow H$, $ad_0 \leftarrow AD(H)$
2: Construct an adapted parametric flow network $N_R$ and $\lambda = AD(H_0)$
3: obtain $H_0'$ from min s-t cut in $N_R$
4: while $l(ad_0, H_0 \leftarrow H_0')) \neq 0$ do
5:     $ad_0 \leftarrow AD(H_0), \lambda = ad_0$
6:     obtain $H_0'$ from min s-t cut in $N_R$
7: end while
8: return $H_0$
```

where $w(e(u,v))$ is the weight of edge (u,v) and $A(u)$ and $A(u)$ are the set of attributes of vertex u and v respectively.

$$\text{context weighted density: } CW(H) = \frac{\sum_{\Delta \in Tri(\Delta)} w(\Delta) + \sum_{e \in E(H)} w(e)}{|V(H)|} \tag{6}$$

Algorithm 1 shows how to find required skills using a tailored parametric preflow algorithm. It starts by considering the whole input subgraph H as a candidate team. The candidate team is the group of members who satisfies the query criteria. In the line 2, We construct a parametric flow network based on the steps explained in the part below. Then, we use the stop condition in line 3 to check whether the subgraph H itself is a candidate team or not. If not we generate a better solution by solving sub problem $l(ad_0, H_0 \leftarrow H_0')$, is defined as below [3],

$$l(ad_0, H_0 \leftarrow H_0') = \sum_{\Delta \in Tri(H_0')} w(\Delta) + \sum_{e \in E(H_0')} w(e) - ad_0 \times (|V(H_0')|) \tag{7}$$

Algorithm 1 considers the progressively modified $ad(H_0)$ as a parameterized capacity in N_R. The overall structure of the algorithm is similar to optimization algorithm, i.e., it continuously generates H_0 with higher context weighted density until reaching the stop condition. During each iteration, internally the algorithm maintains preflow labels via updating the labels computed from the previous iteration. In order to compute H_0', preflow value and some edge capacities are updated according to H_0 generated in the previous iteration. The improved solution gets generated repeatedly until the stop condition is met, i.e., a candidate team is found.

4.4 Complexity Analysis

Assume $|V| = n$ and $|E| = m$. In first step of Geo-social team Formation, it takes $O(m^{1.5})$ time to compute structural similarity. As a result, on each machine M_i extracting all the core nodes from G_i takes $O(m^{1.5})$ time. Consequently finding all dissimilar edges of E_i requires $O(\frac{m}{|M|})$ time. The skills checking complexity can be bounded by $O(|V(cluster)^3|)$, making use of the maximum-flow

Dataset	# of Nodes	# of Edges	Ave-Deg
Gowalla [5]	196,591	950,327	9.177
Dianping [2]	2,673,970	1,922,977	12.184
Orkut-2007 [25]	3,072,626	34,370,166	76.277
Ljournal-2008 [14]	5,363,260	79,023,142	14.734
Twitter-2010 [11]	41,652,098	1,468,365,182	35.253

Table 1. Statistics of real-world datasets.

algorithm. However, providing parametric-network flow help us to solve this in a time complexity of solving one min-cut problem.

5 Experimental Analysis

We conducted experiments to demonstrate the efficiency and effectiveness of our framework. From the efficiency point of view, we show that the Geo-Social team formation model is faster than the state-of-the-art algorithms on large graphs. The proposed framework finds a resulting team with specified features in a Geo-social network having 1.5 billion edges within $8s$. Furthermore, for demonstrating the effectiveness of DSCAN-GSTF, two illustrative queries are analyzed on various real datasets and various metrics based on spatial and social cohesiveness.

Dataset: Table 1 describes the statistical information of five real-world GeoSNs with ground-truth clusters use to evaluate our framework.

5.1 Experiment Setup

We compare our framework, DSCAN-GSTF, with state-of-the-art models MKCSSG [4] and geo-social group queries model (GSGQ) [28]. The MKCSSG model satisfies minimum keyword, contribution capacity, as well as social and spatial constraints. The GSGQ did not consider the required number of experts for each skill. Therefore, we change the GSGQ and add the skill constraint to the team's required skills query such that the skills attributes of the members in the resulting team should cover all required skills.

All the above models are implemented in python using NetworkX, Panda, Tensorflow, Numpy, and pyWebGraph libraries. For the distributed and multiple processing in DSCAN-GSTF, we used MPI. All the experiments are performed on a computer cluster of 16 machines with an interconnecting speed of 9.6GB/s running GNU/Ubuntu Linux 64-bit. Furthermore, each machine's specifications were Intel Xeon E5-2665 64-bit CPU and 256 GB of RAM (8 GB / core). Moreover, MKCSSG and GSGQ are implemented on one machine since they are not distributed algorithms. Each model is executed 20 times, and the average score is recorded.

5.2 Effectiveness Evaluation

To show the effectiveness of the Geo-Social team formation framework, we analyzed two representative queries on the Gowalla dataset.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_44

Model	SC	GD	ED	Query
GSGQ	0.14	0.71	0.51	Query 1: Food
MKCSSG	0.18	0.41	0.67	
DSCAN-GSTF	0.21	0.23	0.74	
GSGQ	0.56	0.54	0.38	Query 2: Music Band
MKCSSG	0.67	0.28	0.63	
DSCAN-GSTF	0.56	0.23	0.81	
GSGQ	0.27	0.62	0.43	Query 3: Board Game
MKCSSG	0.43	0.32	0.52	
DSCAN-GSTF	0.42	0.25	0.68	

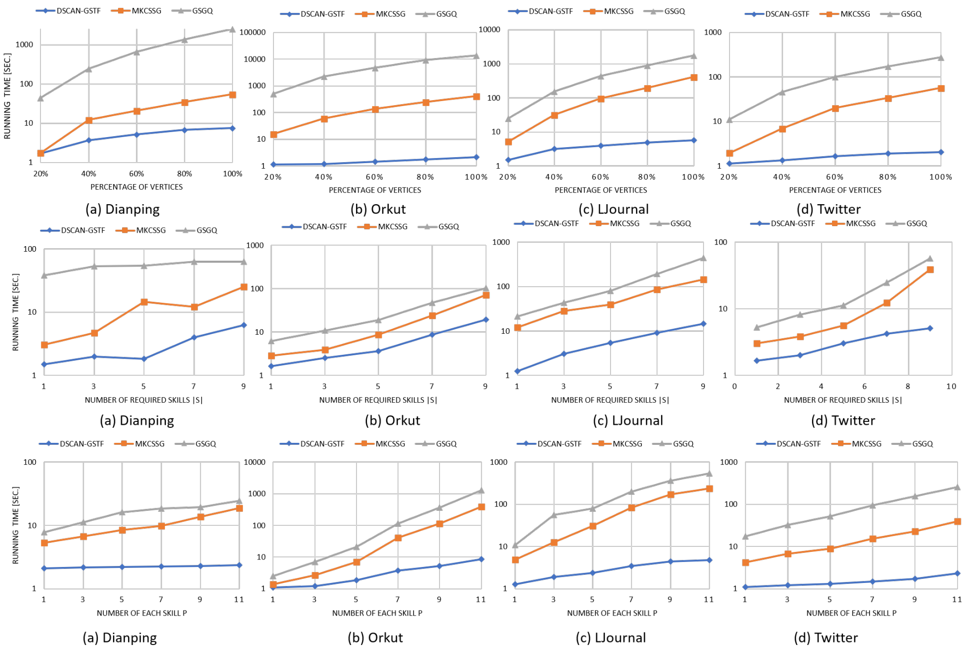
Table 2. Effectiveness evaluation

Query 1: Parameters are set as follow: location $\nabla = (36.11, -115.13)$, Skills $S = \{salad, chicken, beef, BBQ\}$, Number of each skills $E = \{10, 10, 10, 10\}$, Contribution capacity $c = 4$. This query can be used to find fans of BBQ party around Las Vegas. We assume that the tweets of each user is their favorite dish. We set $\epsilon = 0.5$ and $\mu = 10$ for the first query on Geo-Social team formation framework.
Query 2: intends to create a music band around Las Vegas. Query 2 parameters are set as follows: location $\nabla = (36.11, -115.13)$, Skills $S = \{guitar, piano, violin\}$, Number of each skills $E = \{2, 1, 2\}$, Contribution capacity $c = 1$. We set $\epsilon = 0.5$ and $\mu = 10$ for the second query on Geo-Social team formation model.
Query 3: intends to create a board game groups. Query 3 parameters are set as follows: location $\nabla = (36.11, -115.13)$, Skills $S = \{chess, backgammon, monopoly\}$, Number of each skills $E = \{2, 8, 10\}$, Contribution capacity $c = 2$. We set $\epsilon = 0.4$ and $\mu = 9$ for the second query on Geo-Social team formation model.

Evaluation Metrics Here we define the evaluation metrics use to compare the performance of state-of-the-art methods with DSCAN-GSTF.
Spatial Closeness (SC): The spatial cohesiveness is to show how closely the team members are located to ∇. Our framework uses the spatial distance of one random member of the result team B to the query location $\nabla . SC_{\nabla,B} = \{(Euclidean\_Distance(\nabla, u)), u \in V(B)\}$
Graph diameter (GD): It calculates the topological length or extent of a graph by counting the number of edges in the shortest path between the most distant vertices. In other words, graph diameter indicates the **social closeness** of the team. $GD_B = max\{ShortestPath(v, w))|(v, w) \in V(B)\}$
Edge density (ED): We consider another parameter ED to show the social cohesiveness.$ED_B = |E(B)|/|V(B)|$, where $E(B)$ is the number of edges in team B and $V(B)$ is the number of vertices in team B.

The comparison results of analyzed metrics are presented in Table **??**. The results are normalized to a value between 0 and 1. Results with a lower score are better except for Edge Density (ED). Overall speaking, we can see DSCAN-GSTF has outperformed in spatial and social cohesiveness in both queries. However, the spatial distance is not significantly better, but the social cohesiveness shows excellent improvement in both queries. Furthermore, applying graph structural communities in DSCAN-GSTF framework improves the social Cohesion

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_44



(a) Dianping (b) Orkut (c) LJournal (d) Twitter

(a) Dianping (b) Orkut (c) LJournal (d) Twitter

(a) Dianping (b) Orkut (c) LJournal (d) Twitter

Fig. 4. (a) The first raw is to compare the efficiency evaluation based on percentage of vertices. (b)The middle raw is to compare the efficiency evaluation on various number of required skills. (c) The last raw is to compare the efficiency evaluation on various requirement on expert set R.

and indicates teams with much less graph diameter than GSGQ and MKCSSG, which utilize the minimum degree and c-truss constraint, respectively.

5.3 Efficiency Evaluation

To compare the performance of various models, we use the running time of the queries. We compare the efficiency of GSGQ and MKCSSG with our proposed model, DSCAN-GSTF. Our experiments uses various parameter settings for a query: percentage of vertices, sets of required skills $|S| = \{1, 3, 5, 7, 9\}$, and the minimum number of each skill $E = \{1, 3, 5, 7, 9\}$. We set the default value of both $|S|$ and E to 3. The location for each query is created randomly. We select reasonable values for ϵ and μ based on each dataset. In Dianping dataset $\epsilon = 0.3$ and $\mu = 2$, in the Orkut dataset $\epsilon = 0.5$ and $\mu = 5$, in the LJournal dataset $\epsilon = 0.6$ and $\mu = 5$, and finally in the Twitter dataset $\epsilon = 0.5$ and $\mu = 6$. When a parameter is changing for evaluation, other parameter values are set to their default value.

We divide each dataset into various percentages to evaluate the scalability of proposed model. The result is presented in Figure 4 (a) for different datasets while comparing various methods. In general, our DSCAN-GSTF is much more scalable compared to other methods on different datasets. That is because of the methodology and substantially the properties of DSCAN-GSTF which can limit the search space in quicker time while preserving optimum results.

Figure 4 (b) shows the running time when the number of required skills increases for different datasets; as the required skills increase, the running time for all methods increases. However, this increase is slower in DSCAN-GSTF because checking existing skills on each cluster using the attached summation list of skills has constant time complexity. When $|S|$ is small, the GSGQ requires comparatively high running time to find optimum results. However, when the $|S|$ grows, it provides a result in half a minute. In Figure 4 (c), changing the number of required skills E on each dataset using various models is presented. Again, for all the datasets, the running time increase as the required number of skills is increased. Nevertheless, because of distributed environment in DSCAN-GSTF, the increasing running time is on a slow increasing slope.

6 Conclusions

This paper has explored the Geo-Social team Formation framework and proposed a new model DSCAN-GSTF. In this, we incorporated various criteria to replicate the real-world scenario and exploited DSCAN for the efficient process of large networks. The DSCAN-GSTF introduced a novel polynomial algorithm based on a parametric flow algorithm to identify the successful team members for impromptu activities from GeoSNs. We compared our proposed DSCAN-GSTF model with the state-of-the-art methods, MKCSSG and GSGQ. Extensive experiments were conducted to examine the efficiency and effectiveness of the proposed model on four real-world datasets and recorded the running time under various system settings. Overall, our proposed model generated the output faster than the state-of-the-art methods. As for future work, we plan to extend DSCAN-GSTF to incorporate more sophisticated queries.

References

1. Armenatzoglou, N., Papadopoulos, S., Papadias, D.: A general framework for geo-social query processing. Proceedings of the VLDB Endowment pp. 913–924 (2013)
2. Bu, Jiahao, L.S.Y., Wang, J., Zhang, F., Wu, W.: Asap: A chinese review dataset towards aspect category sentiment analysis and rating prediction (2021)
3. Chen, L.: Efficient cohesive subgraph search in big attributed graph data (2018)
4. Chen, L., Liu, C., Zhou, R., Xu, J., Yu, J.X., Li, J.: Finding effective geo-social group for impromptu activities with diverse demands. In: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020)
5. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (2011)
6. Cliquet, G., Baray, J.: Location-based Marketing: Geomarketing and Geolocation. John Wiley & Sons (2020)
7. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to algorithms. MIT press (2009)
8. Gallo, G., Grigoriadis, M.D., Tarjan, R.E.: A fast parametric maximum flow algorithm and applications. SIAM Journal on Computing **18**(1), 30–55 (1989)

9. Huang, Q., Liu, Y.: On geo-social network services. In: 2009 17th International Conference on Geoinformatics. pp. 1–6. Ieee (2009)
10. Kargar, M., An, A., Zihayat, M.: Efficient bi-objective team formation in social networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (2012)
11. Kwak, Haewoon, C.H.S.: What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web (2010)
12. Lappas, T., Liu, K., Terzi, E.: Finding a team of experts in social networks. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 467–476 (2009)
13. Liu, W., Sun, W., Chen, C., Huang, Y., Jing, Y., Chen, K.: Circle of friend query in geo-social networks. In: International Conference on Database Systems for Advanced Applications. pp. 126–137. Springer (2012)
14. Rossi, R., Ahmed, N.: The network data repository with interactive graph analytics and visualization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 29 (2015)
15. Selvarajah, K.: Investigation of team formation in dynamic social networks (2020)
16. Selvarajah, K., Zadeh, P.M., Kargar, M., Kobti, Z.: Identifying a team of experts in social networks using a cultural algorithm. Procedia Computer Science (2019)
17. Selvarajah, K., Zadeh, P.M., Kobti, Z., Kargar, M., Ishraque, M.T., Pfaff, K.: Team formation in community-based palliative care. In: Innovations in Intelligent Systems and Applications (2018)
18. Shen, C.Y., Yang, D.N., Huang, L.H., Lee, W.C., Chen, M.S.: Socio-spatial group queries for impromptu activity planning. IEEE Transactions on Knowledge and Data Engineering **28**(1), 196–210 (2015)
19. Shen, C.Y., Yang, D.N., Lee, W.C., Chen, M.S.: Spatial-proximity optimization for rapid task group deployment. ACM Transactions on Knowledge Discovery from Data (TKDD) **10**(4), 1–36 (2016)
20. Shiokawa, H., Takahashi, T.: Dscan: Distributed structural graph clustering for billion-edge graphs. In: International Conference on Database and Expert Systems Applications. pp. 38–54. Springer (2020)
21. Sohail, A., Cheema, M.A., Taniar, D.: Geo-social temporal top-k queries in location-based social networks. In: Australasian Database Conference (2020)
22. Tsourakakis, C.: The k-clique densest subgraph problem. In: Proceedings of the 24th international conference on world wide web. pp. 1122–1132 (2015)
23. Valverde-Rebaza, J.C., Roche, M., Poncelet, P., de Andrade Lopes, A.: The role of location and social strength for friendship prediction in location-based social networks. Information Processing & Management **54**(4), 475–489 (2018)
24. Xu, X., Yuruk, N., Feng, Z., Schweiger, T.A.: Scan: a structural clustering algorithm for networks. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 824–833 (2007)
25. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. Knowledge and Information Systems **42**(1), 181–213 (2015)
26. Zhao, W., Martha, V., Xu, X.: Pscan: a parallel structural clustering algorithm for big networks in mapreduce. In: International Conference on Advanced Information Networking and Applications (AINA) (2013)
27. Zheng, Y.: Location-based social networks: Users. In: Computing with spatial trajectories, pp. 243–276. Springer (2011)
28. Zhu, Q., Hu, H., Xu, C., Xu, J., Lee, W.C.: Geo-social group queries with minimum acquaintance constraints. The VLDB Journal (2017)