Balancing Authority and Diversity in Influential Node Mining

Yu Zhang
University of Illinois at Urbana-Champaign
yuz9@illinois.edu

ABSTRACT
Diversity is a crucial criterion in many ranking and mining tasks. In this paper, we study how to incorporate node diversity into influence maximization (IM). We consider diversity as a reverse measure of the average similarity between selected nodes, which can be specified using node embedding or community detection results. Our goal is to identify a set of nodes which are simultaneously influential and diverse. Three most commonly used utilities in economics (i.e., Perfect Substitutes, Perfect Complements, and Cobb-Douglas) are proposed to jointly model influence spread and diversity as two factors. We formulate diversified IM as an optimization problem of these utilities, for which we present two approximation algorithms based on non-monotonic submodular maximization and traditional IM respectively. Experimental results show that our diversified IM framework outperforms other natural heuristics, such as embedding and diversified ranking, both in utility maximization and result diversification.

CCS CONCEPTS
• Information systems → Data mining; • Theory of computation → Design and analysis of algorithms;

KEYWORDS
influence; authority; diversity; information networks

1 INTRODUCTION
Many retrieval and mining tasks are concerned with finding the most important or relevant items from a large candidate set. In information network analysis, the task becomes influential node mining. Influence maximization (IM) [24] is one of the solutions proposed for this task. Different from the ranking framework [32], IM focuses on how to select a seed set to activate as many nodes as possible within certain diffusion models.

Although IM has been extensively studied, existing approaches disregard another concern, which is the diversity of selected nodes. In many other tasks, such as information retrieval [8], recommendation [47] and event detection [36], diversity is regarded as a crucial criterion. For example, in document ranking and node ranking, previous approaches tend to balance relevance/authority and diversity in top-ranked items. From the optimization perspective, a weighted sum of the two objectives is maximized [8, 17, 22, 43].

Diversity should be incorporated into the IM framework from two perspectives. (1) Practically, some real applications require the obtained nodes to be not only influential but also diverse. Consider the task of setting up a conference programming committee [30]. The goal is to invite influential researchers from all related areas. Without diversity, the interest of the committee could be biased.

(2) Theoretically, spreading models cannot be 100% accurate in describing real diffusion patterns. To be specific, the parameters are usually set directly instead of learned from data. Even if the parameters are learned, all estimators could have errors, especially for the prediction of weak ties [19] connecting two communities. If all the nodes we select are from one community, we are actually putting all eggs in one basket, which will be risky to achieve wide influence. In contrast, if we choose nodes from different groups or with different features, the solution will become more robust.

In this paper, we study how to balance authority (i.e., the influence spread) and diversity in IM. We define diversity as a reverse measure of the average similarity between selected nodes. According to sociology studies [29, 34], nodes sharing similar properties or types tend to link with each other. Therefore, by incorporating node embedding [20, 33, 39] or community detection [45] techniques, diversity can be specified using structural proximity in the network. To formulate a joint optimization framework of authority and diversity, we would like to go beyond the weighted sum. We adopt three most commonly used two-factor utility functions in economics [44]: Perfect Substitutes, Perfect Complements, and Cobb-Douglas Utility. They describe user’s preference of different choices, and Perfect Substitutes is essentially the weighted sum.

We solve the optimization problem from two directions. In one direction, we prove the three utilities are all submodular (but not always monotonic). This conclusion holds very broadly, so long as the influence function is monotonic and submodular (e.g., Independent Cascade model, Linear Threshold model [24] and many of their variants) and the similarity between two nodes ranges from 0 to 1 (e.g., the neighborhood- and community-level similarities defined in LINE [39] and BigCLAM [45], respectively). Therefore, the RandomGreedy strategy [7] for size-constrained non-monotonic submodular optimization can guarantee a \(\frac{1}{e}\) approximation ratio. In the other direction, we prove that traditional influence maximization, which only focuses on authority, has an impossibility result from the axiomatic perspective.

We conduct extensive experiments on four real-world networks with different choices of spreading models, similarity functions...
and utility functions. Our method consistently outperforms several baselines including node embedding [39], random walks [32] and diversified ranking [22]. Besides, putting the utilities aside, we prove the success of our diversified IM framework from the views of coverage [46], density [30] and concrete cases, indicating that the result is truly diversified, not just maximizing an objective. Moreover, we discuss the practical properties and applicable settings for each utility function.

**Related Work.** Kempe et al. [24] first study Influence Maximization under Independent Cascade and Linear Threshold models. Subsequent efforts following this framework can be divided into two directions. In one direction, researchers focus on how to accelerate the vanilla hill-climbing greedy algorithm [5, 12–14, 27, 40–42]. In the other direction, researchers propose new spreading models and apply them to a wider range of scenarios [10, 11, 18, 23, 25, 28, 37]. Due to space limitation, we do not list all the variants here. One can refer to a recent tutorial [4] for more details. Despite their success, the final cascade size is the only criterion for the IM framework to choose influential nodes. In contrast, our model incorporates diversity as a factor.

Our work is related to result diversification, which is first studied in information retrieval [8]. Subsequent models [1, 2, 6, 9, 17] are all proposed to describe the trade-off between relevance and diversity. Some of these methods show their power in IR, but they hinge on specific choices of similarity functions and cannot be easily generalized to social network scenarios. Others require the dissimilarity function between two elements to form a metric, which is not true in some settings (e.g., the popular node embedding setting discussed below).

To apply diversified ranking into social networks, [46], [30] and [43] adopt random walk based algorithms. He et al. [22] further generalize the framework for arbitrary similarity functions. They require the authority part (or ‘relevance/prestige’ in their papers) to be modular, which does not hold for IM. To the best of our knowledge, [38] is the only previous work exploring diversity in influence maximization. However, their goal is to maximize the entropy of the activation probabilities for each node. This setting is totally different from the original one in IR talking about the diversity of ranked items. Besides, all of the previous trade-off methods between authority and diversity are essentially a weighted sum. In contrast, our model has a wider definition of utility based on economics [44].

## 2 PROBLEM STATEMENT

**Authority.** To capture the dynamics of ads propagation in social networks, two spreading models [24] are commonly studied.

*Independent Cascade (IC) Model.* When a node \( u \) becomes active at time \( t \), it gets a chance to activate each of its inactive neighbor \( v \) at time \( t+1 \), with probability \( p_{uv} \). If \( v \) is not influenced by \( u \), \( u \) cannot make any further attempts in the subsequent rounds.

*Linear Threshold (LT) Model.* Each node \( v \) has a threshold \( \theta_v \sim U[0, 1] \) and each edge \((u, v)\) has a weight \( b_{uv} \). If at time \( t \), we have \( \sum_{u \in v's \\text{active neighbor}} p_{uv} \geq \theta_v \) for an inactive node \( v \), then \( v \) will become active at time \( t+1 \).

In a network \( G = (V, E) \), the campaign spreads from an initially active set of seed nodes \( S_0 = S \) and propagates under certain spreading models as the time runs forward. For IM, we define the **authority** as the expected cascade size (i.e., the number of nodes activated at last), or

\[
\sigma(S) = \mathbb{E}[|S_N|].
\]

Although we focus on IC and LT models in this paper, our results apply to any spreading model with a monotonic and submodular authority function, including many variants of IM models (to name a few, [18], [11] and [28]).

**Diversity.** As we mentioned, besides authority, diversity is another factor whose value should be considered due to the IM framework itself as well as the goal of real applications. In information retrieval, to diversify the ranking result, the relevance score is penalized by the sum of pairwise similarities between the retrieved documents [22, 43]. Here we adopt an analogous definition of **diversity** on the network:

\[
d(S) = 1 - \sum_{u, v \in S, u \neq v} \text{Sim}(u, v) / Z,
\]

where \( \text{Sim}(u, v) \) denotes the similarity between nodes \( u \) and \( v \). \( Z \) is a constant normalization factor to guarantee that \( d(\cdot) \) is non-negative.

The form of Sim(·, ·) can be defined from various perspectives (e.g., semantics, types, etc.). In general, sociologists believe network structures already reflect the similarity between nodes. For example, people sharing hobbies have a higher chance of becoming friends [29]; researchers with common interests are more likely to publish together [34]. Therefore, in this work, we specify diversity using the following two popular settings based on network structures.

*Node Embedding.* For each node \( u \), a low-dimensional vector \( \tilde{e}_u \) is learned to preserve the proximity in the original network. There are several well-known node embedding algorithms [20, 33, 39], among which LINE [39] explicitly defines the proximity between two nodes. In (first-order) LINE, the probability of edge \((u, v)\) is defined as

\[
p_L(u, v) = \frac{1}{1 + \exp(-\tilde{e}_u^T \tilde{e}_v)} = \text{Sigmoid}(\tilde{e}_u^T \tilde{e}_v).
\]

*Community Detection.* For overlapping community detection, BigCLAM [45] uses a vector \( F_u = [F_{u1}, ..., F_{uC}] \) to represent node \( u \), where \( F_u^c \) is the probability that \( u \) belongs to community \( c \). In [45], the probability of edge \((u, v)\) is defined as \( p_C(u, v) = 1 - \exp(-F_u^c \cdot F_v^c) \). The same formula is adopted for disjoint community detection (in which \( F_u \) becomes a “one-hot” vector), we will have

\[
p_C(u, v) = (1 - \frac{1}{e}) \cdot 1_{c_u = c_v},
\]

where \( c_u \) is the community index of \( u \).

We can define \( \text{Sim}(u, v) = p_L(u, v) \) or \( p_C(u, v) \) in order to capture the neighborhood- or community-level proximity.

Note that when \( Z = k(k-1) \) and \( |S| = k \), we have

\[
d(S) = \frac{1}{k(k-1)} \sum_{u, v \in S, u \neq v} \left( 1 - \text{Sim}(u, v) \right) = \frac{1}{k(k-1)} \sum_{u, v \in S, u \neq v} \text{Dis}(u, v),
\]

where \( \text{Dis}(u, v) = 1 - \text{Sim}(u, v) \) is the distance between \( u \) and \( v \). In this case, diversity can be explained as the average pairwise distance in \( S \), same as the definition in [6, 17]. We want to mention that our
In economics, the utility function usually has properties that allow one to substitute authority with diversity at a constant rate. (b) Perfect Complements: When authority and diversity are balanced, neither extra authority nor extra diversity can do us any good. (c) Cobb-Douglas: $z = x^a y^b$. The marginal rate of substitution diminishes.

Dis($\cdot$, $\cdot$) need not be a metric. For example, in the node embedding setting, Dis($u$, $v$) = $1 - \text{Sigmoid}(e_u^T e_v)$, and then Dis($u$, $u$) ≠ 0.

Utility. We have already defined authority and diversity. However, how to compose these two factors into our objective function remains a problem. For instance, if one seed set $S_1$ has authority 100 and diversity 0.6 and another $S_2$ has authority 90 and diversity 0.8, which one do we prefer? Here we adopt three commonly used two-factor utility functions in economics [44].

Perfect Substitutes (PS). Two goods are perfect substitutes if the consumer is willing to substitute one good for the other at a constant rate (e.g., for most people, Pepsi and Coke). In mathematics, $f_S(S) = \sigma(S) + c \cdot d(S)$, where $c$ is a non-negative constant multiplier. To make $f_S(\cdot)$ more meaningful, we can modify $c$ to let $c \cdot d(S)$ share the same magnitude with $\sigma(S)$ (e.g., $c = |V|$).

Essentially, Perfect Substitutes utility is a weighted sum of authority and diversity. [17] and [6] also studied this utility function for diversified ranking. However, in their models, Dis($\cdot$, $\cdot$) must form a metric. Without this assumption (e.g., in the node embedding setting), their algorithms do not have approximation guarantees.

Perfect Complements (PC). A nice example is that of left shoes and right shoes. If we have exactly two pairs of shoes, then neither extra left shoes nor extra right shoes can do us a bit of good. When authority and diversity are regarded as perfect complements, we have $f_C(S) = \min\{\sigma(S), c \cdot d(S)\}$, $c \geq 0$.

Cobb-Douglas Utility (CD). In economics, the utility function usually follows the law of diminishing returns: Adding more of one factor, while holding all other constant, will yield lower incremental per-unit returns. Cobb-Douglas utility is commonly used to describe this property: $f_D(S) = \sigma(S)^a \cdot d(S)^b$, $0 < a, b \leq 1$.

With the three types of utilities, we can formally define the Diversified Influence Maximization (DIM) problem.

Definition 1 (DIM). In a network $G = (V, E)$, given a utility function $f \in \{f_S, f_C, f_D\}$, $\max_{|S| \leq k} f(S)$.

Since DIM can be viewed as an extension of IM, its hardness is straightforward.

Theorem 2.1. DIM is NP-hard for PS, PC and CD utilities.

3 ALGORITHMS

3.1 RandomGreedy

Due to the NP-hardness, we focus on finding algorithms with approximation guarantees. As we all know, the well-known hill-climbing greedy algorithm [24], which has a $(1 - 1/e - \epsilon)$ approximation ratio, requires monotonicity and submodularity of the objective. The bad news here is none of the three types of utilities is monotonic. However, the good news is we still have the submodularity.

Theorem 3.1. For any monotonic and submodular $\sigma(\cdot)$ and any $\text{Sim}(\cdot, \cdot) \in [0, 1]$, the three utilities $f_S(S)$, $f_C(S)$ and $f_D(S)$ are all non-negative and submodular.

Theorem 3.1 naturally applies to authority functions based on IC and LT models, and to diversity functions in our node embedding and community detection settings.

According to Theorem 3.1, DIM is essentially a size-constrained non-monotonic submodular maximization problem. The RandomGreedy strategy proposed in [7] can be adopted for this setting. RandomGreedy is a natural generalization of vanilla hill-climbing greedy. Instead of picking the best single node in each iteration, it first finds $k$ nodes with the highest marginal gains and then randomly selects one node from the top-$k$ candidates to add. A formal description of the algorithm is shown in Algorithm 1. We have the following.

Theorem 3.2 [Buchbinder et al. [7]]. Let $f(\cdot)$ be a non-negative submodular function. For the problem $\max_{|S| \leq k} f(S)$, RandomGreedy guarantees $\mathbb{E}[f(S)] \geq \frac{1}{e} \cdot f(S^*)$, where $S^*$ is the optimal solution.

Theorems 3.1 and 3.2 together guarantee a $(1/e - \epsilon)$ approximated solution for DIM with PS, PC and CD utilities. The “$\epsilon$” here comes from the errors in computing $\sigma(\cdot)$. Chen et al. [12, 14] have pointed out the hardness of this computation under IC and LT models, but an arbitrarily small $\epsilon$ can be obtained through Monte Carlo simulation.

Buchbinder et al. [7] also propose a Continuous Double Greedy strategy. A $(1/e + 0.004)$-factor approximation can be achieved if we adopt the two strategies simultaneously and pick a better one. Due to the very slight difference (0.004) as well as the simplicity and efficiency of RandomGreedy, in our experiments, we prefer to just use RandomGreedy.

Acceleration. The time complexity of RandomGreedy is $O(k|V||C|)$, where $C$ is the time to calculate $f(\cdot)$, or $\sigma(\cdot)$ and $d(\cdot)$. Incrementally updating $d(\cdot)$ only requires $O(k)$ time. For $\sigma(\cdot)$, as mentioned above, it can be approximated through Monte Carlo simulation. Suppose we run $M$ trials of simulation, since each iteration takes $O(|V|)$ time, the overall time complexity will be $O(kM|V|^2)$. We adopt two strategies - LazyForward [27] and StaticGreedy [15] to accelerate RandomGreedy. The reasons why they are useful in submodular cases have been explained in the original papers.
If RandomGreedy, which is just slightly worse. In our experiments, we also observe which adds one step to the simple Greedy algorithm for traditional IMGreedy.

Algorithm 1 RandomGreedy($k$, $f$)
1: initialize $S_0 = \emptyset$
2: for $i = 1$ to $k$ do
3:   Let $M_i \subseteq V - S_{i-1}$ be the subset of size $k$ maximizing $\sum_{x \in M_i} f(S_{i-1} \cup \{x\}) - f(S_{i-1})$
4:   Randomly select $u$ from $M_i$
5:   $S_i = S_{i-1} \cup \{u\}$
6: end for
7: output $S_k$

Algorithm 2 IMGreedy($\sigma$, $f$, $k$)
1: initialize $S_0 = \emptyset$
2: for $i = 1$ to $k$ do
3:   select $u = \arg\max_{x \in V \setminus S_{i-1}} (\sigma(S_{i-1} \cup \{x\}) - \sigma(S_{i-1}))$
4:   $S_{i} = S_{i-1} \cup \{u\}$
5: end for
6: $S^* = \arg\max_{S \in (S_\sigma, \emptyset)} f(S)$
7: output $S^*$

3.2 IMGreedy

One baseline that everyone may think about for DIM is to directly do traditional IM on the network. We propose Algorithm 2 (IMGreedy), which adds one step to the simple Greedy algorithm for traditional IM. It first selects $k$ nodes $S_{\sigma}$ to maximize $\sigma(\cdot)$. Then it compares two solutions $S_{\sigma}$ and $\emptyset$, picking the one with a higher utility. Intuitively, neither $S_{\sigma}$ nor $\emptyset$ is related to the choice of $d(\cdot)$, indicating IMGreedy emphasizes too much on authority and ignores diversity. However, the following result can be proved.

**Theorem 3.3.** For PS utility, IMGreedy provides a $\frac{1}{2}(1 - 1/e - \epsilon)$-factor approximation for DIM.

Note that Theorem 3.3 does not hold for PC and CD. In fact, $f_C(\emptyset) = f_D(\emptyset) = 0$. Besides, since the selection of $S_{\sigma}$ ignores $d(\cdot)$, in the worst/adversarial case, $d(\cdot)$ can be “manipulated” to make $d(S_{\sigma}) = 0$, or $f_C(S_{\sigma}) = f_D(S_{\sigma})$. Hence, IMGreedy can perform arbitrarily poorly for PC and CD.

For PS, the approximation ratio of RandomGreedy is $1/e - \epsilon \approx 0.368$, while IMGreedy achieves a $\frac{1}{2}(1 - 1/e - \epsilon) \approx 0.316$ guarantee, which is just slightly worse. In our experiments, we also observe that RandomGreedy and IMGreedy perform quite close for PS, while RandomGreedy beats IMGreedy by a large margin for PC and CD.

3.3 Monotonicity

One may ask whether we can find a monotonic utility other than the current three. Putting the reasonability aside, such a monotonic utility could let us solve DIM in the traditional greedy framework and obtain a $(1 - 1/e - \epsilon)$ approximation. However, we show an impossibility result from the axiomatic perspective [3, 17].

Suppose the utility function is $f = f(S, \sigma, d)$, where $\sigma$ (resp. $d$) can be any non-decreasing (resp. non-increasing) function defined on $V$. We believe $f$ should satisfy the following two axioms.

**Consistency.** If $\sigma(S_1) \geq \sigma(S_2)$ and $d(S_1) \geq d(S_2)$, then $f(S_1) \geq f(S_2)$. Intuitively, if a set has higher authority and higher diversity at the same time, it should have a higher utility as well.

**Strength of Diversity.** $\exists \sigma(\cdot)$ and $d(\cdot)$, where $\sigma(\cdot)$ is constant and $d(\cdot)$ is not constant, such that $f(S, \sigma, d)$ is not constant. Intuitively, $f(\cdot)$ cannot simply follow $\sigma(\cdot)$ and ignore $d(\cdot)$, otherwise it will lose the ability to model diversity.

**Theorem 3.4.** No monotonic utility $f$ satisfies consistency and strength of diversity at the same time.

4 EXPERIMENTS

We aim to answer two questions in the experimental part. First, does our method achieve higher utilities in contrast to other influential node mining algorithms? Second, putting the utilities aside, does our method really diversify the results without reducing influence?

4.1 Utility Maximization

We have several choices of the objective function. At the basic level, we have three kinds of utilities. Besides, we can choose the spreading model in a network, resulting in different $\sigma(\cdot)$. In this section, we study IC and LT models. For IC model, we let the activate probability $puv$ be $1/\deg_u(v)$. For LT model, similarly, we set the weight on each edge $b_{uv}$ to be $1/\deg_u(v)$. Furthermore, we can select different similarity functions to calculate $d(\cdot)$. Here we consider two aforementioned choices: Community and Embedding.

**BigCLAM** [45] is adopted for community detection, and (first-order) LINE [39] is used to generate node embedding vectors. To summarize, for each dataset, we have 12 choices (namely, {Community, Embedding} × {IC, LT} × {PS, PC, CD}).

**Datasets.** We utilize the following two networks:

1. **Epinions** [35] is a who-trust-whom online social network of a general consumer review site Epinions.com. It is a directed network with 75,879 nodes and 508,837 edges, each edge indicating a user trusts another.

2. **NetHEPT** [13] is an academic collaboration network extracted from arXiv.org, where each node represents an author and each edge means the collaboration in one paper. The network has 15,233 nodes and 59,991 undirected edges.

**Algorithms.** We evaluate the performance of the following influential node mining algorithms:

1. **LINE** [39]: This is a node embedding algorithm. A vector is learned by (first-order) LINE to represent each node. Then we adopt $k$-Means to select the most centered node in each cluster. We use LINE instead of DeepWalk [33] or Node2Vec [20] because it explicitly defines the first-order proximity as $\text{Sigmoid}(\beta_1 \theta^T u_v)$, which is more consistent with our model.

2. **PageRank** [32]: This approach selects nodes with the highest PageRank values.

3. **genDnR** [22]: This is a generic diversified ranking algorithm. It can be applied to the setting with an arbitrary (modular) ranking function and an arbitrary similarity function. Here we use PageRank as the ranking function and $(1 - 1/e)1_{\epsilon_1 = \epsilon_2}$ as the similarity function. (Note that we cannot directly set the ranking function to be $\sigma(\cdot)$ since it is not modular.)
(4) IMGreedy: An extension of traditional IM [24] (which performs better or the same). We use LazyForward and StaticGreedy (with 100 snapshots) to accelerate the algorithm.

(5) RandomGreedy: We set $a = b = 1$ for CD, and $c = 0.05|V|$ for PS and PC. $Z = k(k - 1)$ for each $k$. LazyForward and StaticGreedy strategies are also utilized.

Results. Figure 2 shows the utilities of selected nodes on Epinions. Under PC and CD settings, RandomGreedy consistently outperforms other methods by a large margin. For PS, RandomGreedy is still the best, but its advantage against IMGreedy is slight. These observations are consistent with Theorems 3.2 and 3.3.

LINE, PageRank and GenDeR do not achieve high utilities in most cases. Similar to IMGreedy, PageRank only considers authority while ignores diversity. In contrast, LINE has a guarantee of diversity, but the centroid of a cluster in the embedding space may not be an influential one. The dissimilarity function is incorporated into GenDeR to make it perform slightly better than PageRank. However, it is still not satisfying in our cases.

In fact, besides talking about whether an algorithm emphasizes more on authority or diversity, we can judge whether a utility is able to "penalize" seed sets with low diversity. We have seen that the "penalty" of PS is not enough in contrast to PC and CD since IMGreedy achieves similar performance with RandomGreedy under PS settings, both theoretically and practically. This is intuitive if we consider the three functions $x + y$, $\min(x, y)$ and $xy$. If $y$ is very small, $\min(x, y)$ will be penalized the most, less for $xy$, and even less for $x + y$.

4.2 Coverage and Density

Higher utilities are good news, but do not necessarily indicate satisfying results. Putting the utilities aside, can we prove our success in diversification from other perspectives? Following the evaluation metrics in [16, 22, 30, 46], we conduct experiments on an actor professional network.

Dataset. IMDb network is constructed from the Internet Movie Database. Each actor/actress is represented by a node, and the edges between two nodes denote their co-starred movies. Unseen by the algorithms, each actor/actress is also associated with a country. The dataset we use\footnote{www.kaggle.com/caroolzhangdc/imdb-5000-movie-dataset} involves 5,044 movies and 6,271 actors/actresses, generating an undirected network with 15,060 edges.

Evaluation Metrics. For diversified ranking on networks, Mei et al. [38] leverage Density of the induced subgraph by the $k$ nodes as a measure. Formally,

$$
\text{Density}(S) = \frac{\sum_{u \in S} \sum_{v \in S, u \neq v} 1_{(u, v) \in E}}{|S| \times (|S| - 1)}.
$$

Besides, Zhu et al. [46] propose two diversity measures in a particular context of ranking movie stars, i.e., Country Coverage

Figure 2: Utility values on Epinions, with different kinds of similarity functions (Community-based or Embedding-based), spreading models (IC or LT) and utilities (PS, PC or CD).
and Movie Coverage, which are the number of distinct countries and movies associated with the selected actors/actresses.

Previous studies [16, 22, 30, 46] expect that higher coverage and lower density indicate influential and diverse results.

Results. The evaluated algorithms remain the same, but now RandomGreedy with three different utilities can be compared in one plot. The results are shown in Figure 3. There are some other interesting diversified ranking algorithms, such as DivRank [30] and GCD [16]. In [22], the authors have shown the advantage of GenDeR against these earlier methods under the same setting. We omit restated comparisons here.

From our perspective, although the three metrics are all comprehensive measures, Country Coverage and Density focus more on diversity while Movie Coverage cares more about authority. In Figure 3(b), LINE and RandomGreedy (PC) perform evidently worse, while the others are on par with each other. In Figures 3(a) and 3(c), LINE and RandomGreedy (PC) become the best two. Without them, RandomGreedy (CD) gives the most diversified results in both plots. We explain these observations from two perspectives. On one hand, if we would like to diversify the results with little reduce in authority, RandomGreedy (CD) is the best choice. On the other hand, the three utilities all have their applicable cases. PS favors authority the most; PC prefers diversity; CD is the most proper choice for a “balance”. This can also be explained using the constant elasticity of substitution (CES) utility [44], where $f_{CES}(S) = (\sigma(S)^\rho + d(S)^\rho)^{1/\rho}$, $(\rho \leq 1)$. It has three most popular special cases: PS ($\rho = 1$), CD ($\rho \rightarrow 0$) and PC ($\rho \rightarrow -\infty$). We can clearly see that CD sits between PS and PC.

4.3 Case Study

Our case study is a visualization of the nodes selected by the compared algorithms. This part has a two-fold role. We would like to show the results in an intuitive way, as well as to illustrate why the baselines perform not so well as RandomGreedy.

Dataset. Les Misérables [26] consists of 77 nodes and 254 edges, where the nodes are characters showing up in Victor Hugo’s novel and the edge represents the co-appearance relationship.

Results. Figure 4 shows the nodes selected by different algorithms when $k = 4$. For Les Misérables, Newman and Girvan [31] give a community detection result, which divides the network into 11 non-overlapping groups. We adopt their result to paint the network, giving nodes in the same community the same color.

LINE chooses nodes from four different communities. As we mentioned, the centered node of a cluster in the embedding space may not be an influential one. The blue node selected by LINE is an example. Even if the centered one is authoritative, there may be better choices. As we all know, two kinds of nodes tend to be influential in the network [21]. One is opinion leader located at the center of a community, and the other is structural hole user at the boundary between two or more communities. In the blue and khaki communities, several structural hole users are more influential than the centroids since they connect multiple large communities.

PageRank, GenDeR and IMGreedy give the identical mining results for Les Misérables. PageRank and IMGreedy have no consideration of diversification, and they select two khaki nodes, which will definitely affect the diversity. GenDeR, as a weighted sum of PageRank and diversity, does not give enough penalty here.

RandomGreedy (CD)’s result can be viewed as an ensemble of the previous two results. It selects both opinion leaders and structural hole users from distinct communities.

5 CONCLUSION

This paper presents a diversified IM framework. Three well-known utilities in economics are innovatively proposed to model authority and diversity as two factors, where diversity is specified as the reverse measure of neighborhood- or community-level similarity. Theoretically, we present two approximation algorithms. RandomGreedy guarantees a $(1/e-\epsilon)$-factor approximation for all three utilities. IMGreedy has a $(\frac{1}{2}(1-1/e-\epsilon))$ approximation ratio for PS, but can be arbitrarily bad for PC and CD. Practically, we conduct experiments on four datasets. The results validate our theoretical claims and demonstrates the power of DIM from the views of utility, coverage, density and concrete cases. Moreover, we discuss the practical properties and applicable settings for each utility function.

There are several open questions that present themselves in light of these results. First, it would be interesting to determine whether the approximation guarantee can be improved if we go beyond submodular optimization. Second, more effort is needed to explore a more comprehensive metric jointly evaluating authority (or relevance) and diversity.

REFERENCES

[1] Z. Abbassi, V. S. Mirrokni, and M. Thakur. Diversity maximization under matroid constraints. In KDD’13, pages 32–40, 2013.
[2] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In WSDM’09, pages 5–14, 2009.
[3] K. Arrow. Social choice and individual values. Cowles Foundation, 1951.
[4] C. Aslay, L. V. Lakshmanan, W. Lu, and X. Xiao. Influence maximization in online social networks. In WSDM’18, pages 775–776, 2018.
[5] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier. Maximizing social influence in nearly optimal time. In SOODA’14, pages 946–957, 2014.
[6] A. Borodin, H. C. Lee, and Y. Ye. Max-sum diversification, monotone submodular functions and dynamic updates. In PODS’12, pages 155–166, 2012.
[7] N. Buchbinder, M. Feldman, J. S. Naor, and R. Schwartz. Submodular maximization with cardinality constraints. In SOODA’14, pages 1433–1452, 2014.
[8] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In SIGIR’98, pages 335–336, 1998.
[9] M. Ceccarello, A. Pietracaprina, and G. Pucci. Fast coreset-based diversity maximization under matroid constraints. In WSDM’18, pages 81–89, 2018.
[10] W. Chen, T. Lin, Z. Tan, M. Zhao, and X. Zhou. Robust influence maximization. In KDD’16, pages 795–804, 2016.
[11] W. Chen, W. Lu, and N. Zhang. Time-critical influence maximization in social networks with time-delayed diffusion process. In AAAI’12, pages 1–5, 2012.
[12] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In KDD’10, pages 1029–1038, 2010.
[13] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In KDD’09, pages 199–208, 2009.
[14] W. Chen, Y. Yuan, and L. Zhang. Scalable influence maximization in social networks under the linear threshold model. In ICDM’10, pages 88–97, 2010.
[15] S. Cheng, H. Shen, J. Huang, G. Zhang, and X. Cheng. Stategreedy: solving the scalability-accuracy dilemma in influence maximization. In CIKM’13, pages 509–518, 2013.
[16] A. Dubey, S. Chakraborti, and C. Bhattacharyya. Diversity in ranking via resistive graph centers. In KDD’14, pages 78–86, 2014.
[17] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In WWWW’09, pages 381–390, 2009.
[18] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In KDD’10, pages 1019–1028, 2010.
Figure 3: Evaluations of result diversity on IMDB. The (Embedding, IC) setting is used. Comprehensively, RandomGreedy (CD) performs the best in diversifying the results without reducing authority.

Figure 4: Visualization of selected nodes on Les Misérables (k = 4). Nodes enlarged are the selected ones. Community structures (given in [31]) are represented by different colors. PageRank, GenDeR and IMGreedy give the identical results. The (Embedding, IC) setting is used, under which RandomGreedy does not know the community information.