Targeted Influence with Community and Gender-Aware Seeding

Styczen, Maciej; Chen, Bing Jyue; Teng, Ya Wen; Pignolet, Yvonne Anne; Chen, Lydia; Yang, De Nian

Citation (APA)
Styczen, M., Chen, B. J., Teng, Y. W., Pignolet, Y. A., Chen, L., & Yang, D. N. (2022). Targeted Influence with Community and Gender-Aware Seeding. In *CIKM 2022 - Proceedings of the 31st ACM International Conference on Information and Knowledge Management* (pp. 4515-4519). (International Conference on Information and Knowledge Management, Proceedings). Association for Computing Machinery (ACM). https://doi.org/10.1145/3511808.3557708

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Targeted Influence with Community and Gender-Aware Seeding

Maciej Styczen
Swiss Federal Institute of Technology
Lausanne (EPFL)
Lausanne, Switzerland
maciej.styczen@epfl.ch

Bing-Jyue Chen
Institute of Information Science
Academia Sinica, Taiwan
chenbingjyue@iis.sinica.edu.tw

Ya-Wen Teng
Institute of Information Science
Academia Sinica, Taiwan
ywteng@iis.sinica.edu.tw

Yvonne-Anne Pignolet
The Dfinity Foundation
Switzerland
yvonneanne@dfinity.org

Lydia Chen
Department of Computer Science
Delft University of Technology
Netherlands
Y.Chen-10@tudelft.nl

De-Nian Yang
Institute of Information Science
Research Center for Information
Technology Innovation
Academia Sinica, Taiwan
dnyang@iis.sinica.edu.tw

ABSTRACT

When spreading information over social networks, seeding algorithms selecting users to start the dissemination play a crucial role. The majority of existing seeding algorithms focus solely on maximizing the total number of reached nodes, overlooking the issue of group fairness, in particular, gender imbalance. To tackle the challenge of maximizing information spread on certain target groups, e.g., females, we introduce the concept of the *community and gender-aware potential* of users. We first show that the network's community structure is closely related to the gender distribution. Then, we propose an algorithm that leverages the information about community structure and its gender potential to iteratively modify a seed set such that the information spread on the target group meets the target ratio. Finally, we validate the algorithm by performing experiments on synthetic and real-world datasets. Our results show that the proposed seeding algorithm achieves not only the target ratio but also the highest information spread, compared to the state-of-the-art gender-aware seeding algorithm.

CCS CONCEPTS

• **Theory of computation → Social networks**; • **Social and professional topics → Gender**.

KEYWORDS

social networks, influence maximization, fairness

ACM Reference Format:
Maciej Styczen, Bing-Jyue Chen, Ya-Wen Teng, Yvonne-Anne Pignolet, Lydia Chen, and De-Nian Yang. 2022. Targeted Influence with Community and Gender-Aware Seeding . In *Proceedings of the 31st ACM International*

1 INTRODUCTION

The influence maximization (IM) problem, which chooses a subset of users in the social network as seeds to maximize the number of influenced users [2, 3, 8, 11, 12, 14], has considerable applications, such as viral marketing, political campaigns, and so on. In [8], it has been proved that IM is NP-hard and a greedy algorithm can generate $(1 - \frac{1}{e})$-approximation solutions. However, traditional IM only focuses on the influence spread and neglects the disparity in underrepresented groups (URGs). For example, when the government wants to disseminate a piece of information about financial aid, a good seed group which maximizes the number of influenced individuals but does not take into account if struggling individuals learn about it, may not reach the population which needs it most.

Extensive research [5, 9, 19, 20, 24] has investigated numerous IM problem, which aim to provide fairness and deal with the disparity under various definitions, such as maximizing the influence on URGs while ensuring a minimum spread on the whole network [5], maximizing the minimum spread among all groups [24], balancing the numbers of seeds from different groups [19], and ensuring the ratios of influenced users in respective groups to be the same [9, 20]. Recently, the Disparity Influence Maximization (DIM) problem was proposed [22] to further focus on promoting URG to reach a target ratio, benefiting real-world applications with diverse needs. However, the Disparity Seeding algorithm [22] proposed to solve DIM is based on the ranking mechanism, neglecting the submodular property of IM and in turn resulting in a limited influence spread.

On the other hand, community-based approaches [4, 7, 10, 16, 18] have shown great strengths in solving IM in terms of both effectiveness and efficiency, since they can narrow down the possible seed candidates. Although the above approaches provide promising solutions to traditional IM, none of them has considered fairness. However, as indicated in previous studies [1, 20, 21], human social networks usually exhibit homophily, a tendency to favor interactions with similar individuals. For example, females are more likely to interact with females to form communities. Hence, an essential question is how to exploit such communities to improve seed selection so that a specified target gender ratio can be achieved. In this

paper, we use the relationships between communities and genders to design a community-based approach to solve DIM.

We start by analyzing a dataset from *Facebook*, where users have three modes of interactions, i.e., likes, comments, and tags. Based on these multi-type interactions, we apply a variant of the Leiden algorithm [23] to discover communities and classify them as *male-dominant, female-dominant, and evenly-distributed communities* depending on their majority gender. Our key findings are (a) the community structure is more crucial than the gender for the homophily phenomenon and (b) evenly-distributed communities are usually more influential than the other types of communities.

Leveraging the observations from our community analysis, we propose *Community and Gender-Aware Seeding (CGAS)* to improve various seeding approaches that are unaware of genders for solving DIM. CGAS introduces the *swapping* mechanism to refine the seed group so that the gender distribution of influenced users satisfies the target ratio specified by DIM. To evaluate the influence of a user on different genders, CGAS proposes a novel measure, the *Gender-aware Potential Influence (GPI)*. Specifically, to gain more influenced users of the target gender, CGAS replaces the seeds having a small GPI on the target gender with users having a large GPI on the other gender while retaining the total influence spread. Note that CGAS offers great flexibility since it is applicable to refine the seed groups discovered by various seeding algorithms. Our experiments on the *Facebook* dataset demonstrate that CGAS can achieve the specified target gender ratio and outperforms Disparity Seeding [22] in terms of the influence spread. Our contributions include:

- We are the first to study the affinity between communities and genders in disparity influence maximization.
- We develop a novel community-gender aware seeding algorithm, CGAS which iteratively refines the seeds based on a new metric, the Gender-aware Potential Influence (GPI).
- The experimental results show that CGAS achieves more than two times the influence spread of Disparity Seeding on the Facebook dataset.

2 COMMUNITY ANALYSIS

2.1 Dataset and community detection

The data was gathered from voluntary senior students at 25 university departments through the Facebook API [22].[1] Specifically, the dataset contains 1870 unique users (765 males and 1,105 females) and around 20 million interactions, each of which records users' genders and the interaction type (i.e., likes, comments, or tags). The period of interactions spans from March 2008 to May 2016.

Following [13], we first discover communities by leveraging a variant of Leiden algorithm [23], which is suitable for multi-type interaction social networks. Next, according to two-third majority known in politics [17], we define a community as a *male-dominant (or female-dominant) community* if the number of males (or females) is more than twice that of females (or males) in this community. Otherwise it is an *evenly-distributed community*. As a result, we obtain 44 communities with an average size of 42.2 (±23.49) members, includiung 21 evenly-distributed (Even), 16 female-dominant (Female), and 7 male-dominant communities (Male).

---

[1]More information about the dataset can be found at https://www2.ios.sinica.edu.tw/sc/en/home2.php.
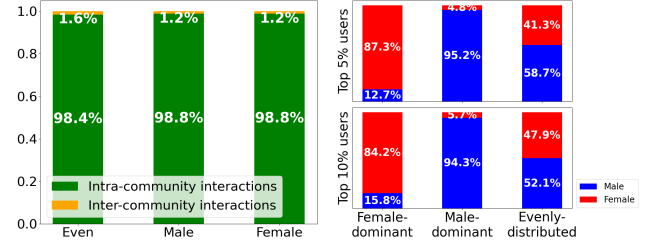
(a) The distribution of inter-/intra-community interactions. (b) The gender distribution of the top-ranked users.

Figure 1: Distributions of user interactions for tagging.
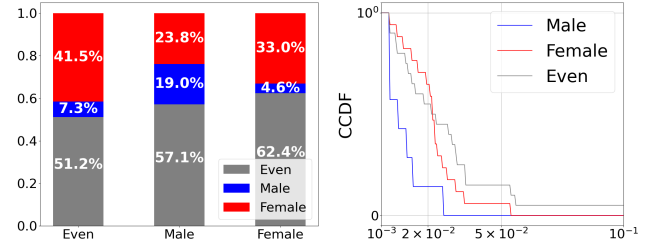
(a) Community interaction. (b) The CCDF of PageRank.

Figure 2: Inter-community interactions for tagging.

2.2 User interaction

We first examine the interaction pattern of users in different types of communities in terms of the type of communities their objects belong to. Due to the space limitation, we follow [22] to focus on the analysis of tags. Figure 1(a) shows the percentage of different types of communities that users in one specific type of community interact with. Take the second bar (i.e., Male) as an example, which shows that the users in male-dominant communities receive and send most tags to users in the same community. Generally, the results demonstrate that more than 98% of users tend to interact with users in the same community regardless the type of communities.

We then analyze whether the gender distribution of top-ranked users (i.e., those who interact with intra-community users the most) in a community complies with the gender that dominates the community. We examine the top 5% and top 10% users as the top-ranked users in Figure 1(b). The results show that the gender distribution of the top-ranked users generally follows the dominant gender of the community. For instance, there are on average 84.2% female users in the top 10% users of a female-dominant community. The observations suggest that with respect to influence in a gender-dominant community, users of the dominant gender are more likely to influence more members of the community.

2.3 Inter-community interaction

Next, we focus on inter-community interactions among different types of communities. Specifically, we first transform the original user-level social network into a community-level network, where nodes are communities and weighted-edges represent the numbers of interactions between members of two different communities. Figure 2(a) shows the distribution of interactions among communities, where the number of interactions (i.e., weights) on each edge is considered. We find that for all types of communities, users have

the least interactions with male-dominant communities. The ratios of users in female-dominant and evenly-distributed communities interacting with those in male-dominant communities (i.e., blue parts) are even lower than that of the male-dominant community on this dataset (i.e., $\frac{7}{44} = 15.9\%$), while more interactions are observed in interactions between male-dominant communities. By contrast, female-dominant and evenly-distributed communities usually tend to interact with each other, i.e., female-dominant communities have the highest ratio to interact with evenly-distributed communities, and vice versa. According to the above observations, different types of communities indeed have diverse patterns in the type of communities they interact with.

Finally, we report the PageRank [15] scores of each community in the community-level graph to examine which type of communities is more influential. Figure 2(a) reveals that evenly-distributed communities are typically the most influential since the complementary cumulative distribution function (CCDF) tail (i.e., the top-ranked communities) of *Even* is greater than those of the other twos. Moreover, as shown in Figure 2(a), evenly-distributed and female-dominant communities interact with each other very frequently. These observations provide evidence that traditional community-based IM approaches that target on the most influential communities may neglect the influence on the underrepresented group.

3 THE DIM PROBLEM AND ALGORITHM

3.1 Problem formulation

In this paper, we aim to solve the Disparity Influence Maximization (DIM) problem, which is formally defined as follows.

Definition 3.1 (DIM [22]). Given a social network $G = (V, E)$, a diffusion model, a parameter k, a target gender ratio ζ, and an error margin e, DIM finds a seed group $S \subseteq V$ with $|S| = k$ to maximize the influence spread with the constraint that the ratio of the target gender in the influenced users is ζ within an error margin e.

3.2 Algorithm

Leveraging the observations from Section 2, we propose *Community and Gender-Aware Seeding (CGAS)* to solve DIM. As seeding algorithms that are unaware of genders fail to achieve the specified influenced ratio of the target gender ζ, CGAS proposes a swapping mechanism to refine the seed group by evaluating users' influence on different genders. To exploit to homophily, CGAS introduces the *Gender-aware Potential Influence (GPI)* by summing up the intra- and inter-community influence on some gender that the neighbors of this gender can make. CGAS exchanges seeds and non-seeds according to their GPI in iterations aiming get closer to ζ while ensuring a large influence spread. In the following, we first formally define GPI and then explain the swapping mechanism in detail.

To evaluate one's influence on different genders, CGAS defines the Gender-aware Potential Influence (GPI) measure. A user v's GPI on a gender $g \in \{M, F\}$ is the sum of the influence on g that v's neighbors of g have. Inspired by the success of community-based algorithms [4, 7, 10, 16, 18], GPI considers the *Gender-aware Community Influence (GCI)* to evaluate neighbors' influence on different genders from the intra- and inter-community aspects.

Formally, we define the GPI of a user v on a gender g as follows.

$$GPI_g(v) = \sum_{w \in N(v) \setminus A, \, w\text{'s gender is } g} b_{vw} \times GCI_g(w), \quad (1)$$

where $GCI_g(w)$ is w's GCI on g (defined later), $N(v)$ is the set of v's neighbors, A is the set of influenced users, and b_{vw} is the weight on the edge between v and w (given by the social network).

Analogously to [16], we evaluate GCI by considering two types of users in communities. One type represents *core users* who only have intra-community interactions, and the other type comprises *boundary users* who interact with users inside and outside the community. For a core user, the GCI on a gender g is based on i) the number of users of g in their community. Moreover, we also add ii) the number of their neighbors of g to account for homophily. For a boundary user, GCI on g is determined by the influence on each community he/she interacts with. Similar to core users, GCI also considers i) the average number of users of g in the communities that a boundary user interacts with and ii) the number of his/her neighbors of g. As observed in Section 2.3, users in different types of communities have diverse patterns in the type of communities they interact with, GCI further takes iii) the number of g-dominant communities he/she interacts with into account. Therefore, the GCI of a user v on a gender g is defined as follows.

$$GCI_g(v) = \begin{cases} U_g(v) + D_g(v) & \text{if } v \text{ is a core user} \\ \alpha \cdot AU_g(v) + D_g(v) + C_g(v) & \text{otherwise} \end{cases}, \quad (2)$$

where $U_g(v)$ is the number of users of g in v's community, $D_g(v)$ is the number of v's neighbors of g, $AU_g(v)$ is the average number of users of g in the communities v interacts with, and $C_g(v)$ is the weighted number of communities v interacts with and the weight of each community is the fraction of users of g. Besides, as users tend to interact more with intra-community users (observed in Section 2.2), $AU_g(v)$ may overestimate v's influence and is adjusted by a parameter $0 < \alpha \leq 1$ reflecting the ratio of inter-community interactions to all interactions.

Equipped with GPI, CGAS applies the swapping mechanism to refine the seed group by iteratively exchanging seeds and non-seeds. Let S_0 denote the initial seed group generated by an arbitrary, possibly non-gender-aware, seeding algorithm. For each iteration $i \geq 1$, CGAS estimates the influence spread of S_{i-1} on each gender to derive the influenced ratio of the target gender r_{i-1}. Based on the comparison between r_{i-1} and ζ (defined in Definition 3.1), CGAS determines the gender that should be mainly influenced in this iteration i, denoted as g_i. Then, CGAS chooses n seeds from S_{i-1} with the lowest GPI_{g_i} as *out-candidates* and n nodes from $V \setminus S_{i-1}$ with the highest GPI_{g_i} as *in-candidates* to form n^2 exchange pairs, where n is a parameter to control the number of exchange pairs.[2] Specifically, CGAS forms an exchange pair by one in-candidate and one out-candidate and estimates the influence spread on each gender considering S_{i-1} is updated by this exchange pair, i.e., removing the out-candidate from the seed group and adding the in-candidate into the seed group. Consequently, CGAS selects the exchange pair with the largest influence spread on g_i to update S_{i-1} as S_i. The

---

[2] A larger n leads to more exchange pairs, which require more search time but provide a higher chance to reach ζ at once.

Table 1: The comparison of influence spread and its female ratios under two community-gender aware seeding.

Specified ratio (ζ)	AN+CGaS			CELF+CGaS			DV		DP	
	Ratio	Spread	#Iter	Ratio	Spread	#Iter	Ratio	Spread	Ratio	Spread
Facebook 0.4	.408	**245.2**	19	.461	337.0	20	.621	223.7	.403	153.6
0.5	.500	228.2	6	.508	**381.8**	4	.621	223.7	.506	112.4
0.6	.607	221.7	3	.594	**413.3**	2	.647	222.4	.595	191.9
0.7	.699	251.9	8	.692	**381.2**	12	.663	219.2	.698	165.2
synthetic 0.3	.305	**41.48**	5	.295	37.50	5	.380	44.07	.305	26.35
0.4	.401	**41.63**	2	.396	39.51	2	.412	45.91	.391	36.67
0.5	.500	**37.01**	5	.502	36.25	5	.448	44.68	.548	33.12
0.6	.603	30.77	7	.604	**32.77**	6	.481	42.73	.592	27.08

swapping mechanism continues until r_i reaches ζ within the error margin e or the maximum iteration $i_{\max}$ is reached, i.e., $i = i_{\max}$.

4 EXPERIMENTS

4.1 Setup

Datasets. We first use the tag interactions on the Facebook dataset (Section 2.1) to evaluate the performance of CGaS. Furthermore, we construct the following synthetic dataset to illustrate our approach on a smaller network. As there are more females than males in the Facebook dataset (i.e., the female ratio is 60%), we apply a modified stochastic block model [6] to generate a synthetic dataset with a community structure that contains more males than females. This dataset consists of four communities with 28, 20, 22, and 30 users, and the corresponding female ratios are 0.8, 0.5, 0.25, and 0.75, respectively. In other words, the synthetic dataset has 60 males and 40 females and the sizes of the communities are in the same range as small communities in the Facebook dataset. To mimic the interaction patterns of social networks like the Facebook dataset, the probabilities of having intra- and inter-community interactions are chosen in ranges [0.7, 0.8] and [0.01, 0.03], respectively.

Baselines. We use two non-gender-aware seeding algorithms and two gender-aware seeding algorithms for comparison. i) Agnostic seeding (AN) [20] selects the top-ranked users as seeds according to the number of users' interactions. ii) CELF [11] greedily selects the users with the maximum marginal gain on the influence spread as seeds. iii) Diversity seeding (DV) [20] selects the top-ranked males and females as seeds according to the number of users' interactions, where the female ratio of the seed group is searched between the specified female ratio ζ and the female ratio of AN's seed group. iv) Disparity seeding (DP) [22] selects the top-ranked males and females as seeds according to the Target HI-Index scores, where the female ratio of the seed group is learned from a scaling function. As AN and CELF are not aware of genders, we exploit CGaS to refine the seed groups derived from them, denoted as AN+CGaS and CELF+CGaS, to demonstrate the applicability of CGaS to different seeding algorithms. Moreover, we compare AN+CGaS and CELF+CGaS with DV and DP to show the superiority of CGaS to maximize the influence spread while approaching ζ.

Metrics and Parameter Settings. We evaluate i) the female ratio of the influenced users and ii) the influence spread by varying the specified female ratios of the influenced users ζ. The sizes k of the seed group of the Facebook and synthetic datasets are set to 25 and 10, respectively, and the error margin for both datasets is set to 0.01. Following [20], the diffusion is simulated through the IC model [8] over 10000 times, where the probability of user v influencing user w

(i.e., the weight b_{vw}) is the number of interactions w receives from v divided by the total number of interactions w receives [14]. For CGaS, α, n and $i_{\max}$ are set as $\frac{1}{3}$, 5 and 20 for both datasets.

4.2 Results

Table 1 lists the female ratio of the influenced users and the influence spread achieved by different algorithms on the Facebook and synthetic datasets. The female ratios with underlines represent the success of achieving ζ within the error margin e, while the spreads with bold are the maximum spread among the algorithms having female ratios with underlines. Note that we vary ζ between [0.4, 0.7] and [0.3, 0.6] for the Facebook and synthetic datasets, respectively, since their dominant genders are different. From the results, we have three observations. (1) CGaS can approximate ζ based on the seed groups obtained by AN and CELF, showing the applicability of CGaS to various seeding algorithms. (2) DV fails to achieve ζ within e and has moderate errors only when ζ is close to the female ratio in the population (i.e., 0.6 and 0.4 in the Facebook and synthetic datasets, respectively). By contrast, AN+CGaS, CELF+CGaS, and DP deviate least from ζ. (3) Although DP has small errors, AN+CGaS and CELF+CGaS always exhibit a much larger influence spread than DP while ensuring ζ is reached, since CGaS exploits the intra- and inter-community influence to evaluate the users' influence. Specifically, CELF+CGaS achieves more than two times the influence spread of DP on the Facebook dataset.

Table 1 further presents the number of iterations required by CGaS to approach ζ. As ζ deviates more from the gender distribution of the population, more iterations are required. Since non-gender-aware seeding algorithms usually yield the gender distribution in the influenced users following that in the population, a larger difference between ζ and the gender distribution of the population results in more efforts to adjust the seed group. For example, CELF+CGaS is unable to reach 0.4 ± 0.01 within $i_{\max} = 20$.

5 CONCLUSION

Motivated by the prevalent community structure and the asymmetric influence of different genders in social networks, we conducted a first-of-its-kind gender-aware community analysis and presented a novel seeding algorithm to promote the information spread on the target group. We studied how the community structure is affected by the gender distribution and showed that many communities have a strong affinity to a single gender. Secondly, we designed a community-gender aware algorithm that achieves a target ratio of information spread by iteratively adjusting the seeding selection based on the community-gender structure. We evaluated the proposed seeding algorithm against the state-of-the-art gender-aware seeding algorithms on both synthetic graphs and a Facebook trace. The results demonstrate that the proposed community-gender aware seeding algorithm achieves the target influence ratio while maximizing the information spread.

REFERENCES

[1] C. Avin, B. Keller, Z. Lotker, C. Mathieu, D. Peleg, and Y.-A. Pignolet. 2015. Homophily and the glass ceiling effect in social networks. In *ACM ITCS*.

[2] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier. 2014. Maximizing social influence in nearly optimal time. In *SIAM SODA*.

[3] X. Chen, Y. Zhao, G. Liu, R. Sun, X. Zhou, and K. Zheng. 2020. Efficient similarity-aware influence maximization in geo-social network. *IEEE TKDE* (2020).

[4] Y.-C. Chen, W.-Y. Zhu, W.-C. Peng, W.-C. Lee, and S.-Y. Lee. 2014. CIM: Community-based influence maximization in social networks. *ACM TIST* (2014).

[5] S. Gershtein, T. Milo, B. Youngmann, and G. Zeevi. 2018. IM balanced: Influence maximization under balance constraints. In *ACM CIKM*.

[6] P. W. Holland, K. B. Laskey, and S. Leinhardt. 1983. Stochastic blockmodels: First steps. *Soc. Netw.* (1983).

[7] H. Huang, H. Shen, Z. Meng, H. Chang, and H. He. 2019. Community-based influence maximization for viral marketing. *APIN* (2019).

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

[9] M. Khajehnejad, A. Asgharian Rezaei, M. Babaei, J. Hoffmann, M. Jalili, and A. Weller. 2020. Adversarial graph embeddings for fair influence maximization over social networks. In *IJCAI*.

[10] S. Kumar, A. Gupta, and I. Khatri. 2022. CSR: A community based spreaders ranking algorithm for influence maximization in social networks. *W3J* (2022).

[11] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. 2007. Cost-effective outbreak detection in networks. In *ACM SIGKDD*.

[12] Y. Li, J. Fan, Y. Wang, and K.-L. Tan. 2018. Influence maximization on social graphs: A survey. *IEEE TKDE* (2018).

[13] P. Mucha, T. Richardson, K. Macon, M. Porter, and J.-P. Onnela. 2010. Community structure in time-dependent, multiscale, and multiplex networks. *Science* (2010).

[14] H. T. Nguyen, M. T. Thai, and T. N. Dinh. 2016. Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks. In *ACM SIGMOD*.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

[16] L. Qiu, W. Jia, J. Yu, X. Fan, and W. Gao. 2019. PHG: A three-phase algorithm for influence maximization based on community structure. *IEEE Access* (2019).

[17] H. M. Robert. 2000. *Robert's Rules of Order* (10 ed.). Da Capo Press, London, England.

[18] J. Shang, S. Zhou, X. Li, L. Liu, and H. Wu. 2017. CoFIM: A community-based framework for influence maximization on large-scale networks. *KBS* (2017).

[19] A.-A. Stoica and A. Chaintreau. 2019. Fairness in social influence maximization. In *WWW*.

[20] A.-A. Stoica, J. X. Y. Han, and A. Chaintreau. 2020. Seeding network influence in biased networks and the benefits of diversity. In *WWW*.

[21] A.-A. Stoica, C. Riederer, and A. Chaintreau. 2018. Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In *WWW*.

[22] Y.-W. Teng, H.-W. Chen, D.-N. Yang, Y.-A. Pignolet, T.-W. Li, and L. Chen. 2021. On influencing the influential: Disparity seeding. In *ACM CIKM*.

[23] V. Traag, L. Waltman, and N. J. van Eck. 2019. From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* (2019).

[24] A. Tsang, B. Wilder, E. Rice, M. Tambe, and Y. Zick. 2019. Group-fairness in influence maximization. In *IJCAI*.