2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) | 978-1-7281-1056-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/ASONAM49781.2020.9381293

Discovering Interesting Subgraphs in Social Media Networks

Subhasis Dasgupta
San Diego Supercomputer Center
University of California San Diego
La Jolla, USA
sudasgupta@ucsd.edu

Amarnath Gupta
San Diego Supercomputer Center
University of California San Diego
La Jolla, USA
a1gupta@ucsd.edu

***Abstract*—Social media data are often modeled as heterogeneous graphs with multiple types of nodes and edges. We present a discovery algorithm that first chooses a "background" graph based on a user's analytical interest and then automatically discovers subgraphs that are structurally and content-wise distinctly different from the background graph. The technique combines the notion of a `group-by` operation on a graph and the notion of subjective interestingness, resulting in an automated discovery of interesting subgraphs. Our experiments on a socio-political database show the effectiveness of our technique.**

***Index Terms*—social network, interesting subgraph discovery, subjective interestingness**

I. Introduction

Social media is often modeled as graphs – graphs where the nodes represent entities (e.g., users. geographic objects), themes (e.g., hashtags), content (e.g., posts, URLs) and so forth, while the edges represent relationships such as "a post *commenting on* another", "a user having a *friendship with* another", "a post *containing* a hashtag" and so forth. For some applications, computationally derived edges are used – for example, hashtag co-occurrence (i.e., the fact that a pair of hashtags has appeared in the same post) is a commonly used derived edge [1]. A typical social media graph has both node properties (e.g., date of a post) and edge properties (e.g., co-occurrence count, the time-interval over which a friendship relationship holds). In addition, a social media graph may have named subgraphs such as user-defined sub-communities (e.g., a Facebook group) which may have their own properties (e.g., the "privacy level" of the group). This paper investigates a technique to discover "interesting subgraphs" from a *Social Media Graph*. We formalize the notion of "interestingness" in Section II. Informally, a subgraph of a social media network is "interesting" if the subgraph has a structure and content that is sufficiently different from the rest of some reference social media network. There are many reasons why a subgraph would be different from the overall tweet graph. Consider the first tweet shown in Table I – the entire tweet has no content, only five mentioned users. When viewed as graph, the tweet nodes have five `mention` edges but content value is `null`. This single tweet is interesting because contentless tweets are statistically rare. Now imagine that a larger tweet graph has small pockets of dense subgraphs consisting of contentless tweets. These subgraphs can be considered "interesting" because they represent a statistically unusual density of statistically unusual content. In contrast, the second tweet in Table I has content discussing the rapper "TI" in a closed group. Even if there are similarly dense subgraphs representing an intense discussion on the theme, it is not necessarily interesting, unless the content of the conversation is very different from the content of the conversation of the graph surrounding it. Thus, the notion of interestingness depends both on the content and structure of the subgraph and can only be interpreted in the context of a reference network as determined by an analyst's need.

II. Interesting Subgraphs of a Social Network

Related Work. The problem of finding interesting subgraphs has been investigated from several different viewpoints. One of the earliest "graph mining" approaches focused on discovering the most frequently occurring subgraphs [2]. A second approach considers interesting subgraphs as a subgraph matching problem [3]. Their general approach is to compute all matching subgraphs that satisfy a user the query and then ranking the results based on the rarity and the likelihood of the associations among entities in the subgraphs. A third approach [4] uses the notion of "subjective interestingness" which roughly corresponds to finding subgraphs whose connectivity properties (e.g., the average degree of a vertices) are distinctly different from an "expected" *background* graph. This approach uses a constrained optimization problem that maximizes an objective function over the *information content* and the *description length* of the desired subgraph pattern. **Our Approach.** We assume that the social media is represented by a social media graph (more generally, any heterogeneous network) G_0. We initiate the discovery process by a user-specified query Q to specify an initial subnetwork $G' = Q(G_0)$, called the *initial background graph* over which the discovery process is conducted. Over G' we discover subgraphs $S_i \subset G'$ whose content and structure are distinctly different that of G'. However, unlike previous approaches, we apply a generate-and-test paradigm for discovery. The generate-step (Section III-A) uses a graph cube like [5] technique to generate candidate subgraphs that might be interesting and the test-step (Section III-B) computes if (a) the candidate is sufficiently distinct from the G', and (b) the collection of candidates are sufficiently

	Interesting Text from Tweets	Why These Tweets are Interesting
1	@rooseveltinst @Justice4ADOS @SandyDarity @IrstenKMullen @MusicNegrito	Creating a strongly connected network by mentioning only users like a robot.
2	@noirdosser @chelleter_d @SandyDarity @quantumblackne2 @Tip @KeishaBottoms @esglaude I think TI is fake shook..... typical move celebrities play.	Creating a close issue centric network by adding known and focused users.
3	@princss6 @DerrickNAACP I agree. At this critical juncture when the natl attention is on injustice to #ADOS he is "all black lives" mattering our justice claim. This makes no sense. #ResignDerrick	While the content is simple, this tweet bridges two different dense subnetworks by co-mentioning two popular users from these two networks.
4	@Hub_Libertarian @davidenrich @realDonaldTrump @DeutscheBank Love how ignored the facts about Supreme Court decisions...lol. 9-0, the most common decision is facts you can't ignore.	Tweets like this are not interesting. They create a focused but broad network by mentioning all related users, some of whom are very popular.
5	@grey_geena @obiora_odi @KHiveQueenBee @livemusic4me @Cat_MarqueeLV @Unknwnstuntman @ElMcClelland @annableigh @thatboybesangin @fourgunfire @moshimisen @sheanabana @twobesure @Alysson @NancyTabak @JoeBiden We have no choice but to let it play out, however, white folks out her writing letters to the manager and equating life long Black public servants to "cosmetics" and "tokens", sooo yeah my trust in "the process" is minimal, right about now.	Creating a broad network by mentioning as many users as possible.
6	@KBULTRA0 @KamalaHarris Tomorrow I will conduct myself the way an old Italian Catholic nona in Napoli celebrates Shivaratri ˆthis is the only "resistance" possible. In fact I've already partially ruined it	These types of tweet are interesting because they gain attention by mentioning popular users who are fairly unrelated to the content of the tweet

TABLE I: Some types of tweets that are more "interesting" than others because the network around these tweets show some unusual phenomena (see text for more explanation).

distinct from each other. **Subgraph Interestingness.** For a subgraph S_i to be considered as a candidate, it must satisfy the following conditions. **C1.** S_i must be connected and should satisfy a size threshold θ_n, the minimal number of nodes. **C2.** Let A_{ij} (resp. B_{ik}) be the set of *local* properties of node j (resp. edge k) of subgraph S_i. A property is called "local" if it is not a network property like vertex degree. All nodes (resp. edges) of S_i must satisfy some user-specified predicate ϕ_N (resp. ϕ_E) specified over A_{ij} (resp. B_{ik}). For example, a node predicate might require that all "post" nodes in the subgraph must have a re-post count of at least 300, while an edge predicate may require that all hashtag co-occurrence relationships must have a weight of at least 10. A user defined constraint on the candidate subgraph improves the interpretability of the result. Typical subjective interestingness techniques [4], [6] use only structural features of the network and do not consider attribute-based constraints, which limits their pragmatic utility. **C3.** For each text-valued attribute a of A_{ij}, let $C(a)$ be the collection of the values of a over all nodes of S_i, and $\mathcal{D}(C(a))$ is a textual diversity metric computed over $C(a)$. For S_i to be interesting, it must have at least one attribute a such that $\mathcal{D}(C(a))$ does not have the usual power-law distribution expected in social networks. Zheng et al [7] used *vocabulary diversity* and *topic diversity* as textual diversity measures.

III. The Generate and Test Process

A. Candidate Generation

Initial Query. The candidate generation process starts with an initial query Q to the social network graph. The query is placed against the original social media data without considering their network structure. For example, a query can select all tweets containing the hashtag `#ADOS` starting in 2019. The resulting collection becomes the universe of discourse for interestingness discovery. The initial background graph G' is constructed on the results of this query. **Node Grouping.** Given the graph G', the user specifies a grouping condition over a node set of the graph. The grouping condition may be specified in two ways: (1) Using a Boolean condition over node properties, e.g., "tweet" nodes can be grouped based on `tweetDate` $\wedge$ `favoriteCount` (binned by 100); (2) Using a grouping pattern, e.g., `(:tweet{date})-[:uses]->(:hashtag{text})` states that all "tweet" nodes having the same posting date, together with every distinct hashtag text will be placed in a separate group. Notice that while (1) produces disjoint tweets, (2) produces a "soft" partitioning on the tweets and hashtags due to the many-to-many relationship between tweets and hashtags. In either case, the result is a set of node groups, designated here as N_i. **Graph Construction.** The graph construction phase constructs a subgraph S_i by expanding on the node set N_i. Different expansion rules can be specified, leading to the formation of different graphs. Here we list three rules that we have found fairly useful in practice. **G1.** Identify all the `tweet` nodes in N_i. Construct a *relaxed induced subgraph* of the `tweet`-labeled nodes in N_i. The subgraph is induced because it only uses tweets contained within N_i, and it is *relaxed* because contains all nodes *directly associated* with these tweet nodes, such as author, hashtags, URLs, and mentioned-users. **G2.** Construct a *mention network* from within the tweet nodes in N_i – the mention network initially connects all `tweet` and `user`-labeled nodes. Extend the network by including all nodes *directly associated* with these tweet nodes. **G3.** A third construction relaxes the grouping constraint. We first compute either **G1** or **G2**, and then extend the graph by including the first order neighborhood of mentioned users or hashtags. While this clearly breaks the initial group boundaries, a network thus constructed includes tweets of similar themes (through hashtags) or audience (through mentions). Once these candidate graphs are constructed, they are tested for criterion **C3**. In this paper, we have directly applied the diversity metric proposed in [7].

B. Testing for Relative Interestingness

We compute the interestingness of a subgraph S in reference to a background graph G_b, and consists of a structural as

well as a content component. We first discuss the structural component. To compare a subgraph S_i with the background graph, we first compute a set of network properties P_j (see below) for nodes (or edges) and then compute the frequency distribution $f(P_j(S_i))$ of these properties over all nodes (resp. edges) of (a) subgraphs S_i, and (b) the reference graph (e.g., G'). A distance between $f(P_j(S_i))$ and $f(P_j(G_b))$ is computed using Jensen–Shannon divergence (JSD). In the following, we use $\Delta(f_1, f_2)$ to refer to the JS-divergence of distributions f_1 and f_2. **High-Centrality Nodes:** The testing process starts by identifying the distributions of nodes with high node centrality between the networks. While there is no shortage of centrality measures in the literature, we choose eigenvector centrality [8]. The rationale for this choice follows from earlier studies in [9]–[11], who establish that it represents the true structure of the network more faithfully than other centrality measures. Let the distributions of eigenvector centrality of subgraphs A and B be β_a and β_b respectively, and that of the background graph be β_t, then $\Delta_e(\beta_t, \beta_a) > \theta$ indicates that A is sufficiently structurally distinct from G_b $\Delta_e(\beta_t, \beta_a) > \Delta_e(\beta_t, \beta_b)$ indicates that A contains more influential nodes than B. **Navigability:** The second network feature we consider is *edge betweenness centrality* [8]. Since edge betweenness centrality of edge e measures the proportion of paths that passes through e, a subgraph S with a higher proportion of high-valued edge betweenness centrality implies that S is more *navigable* than G_b or another subgraph S'. Let the distribution of the edge betweenness centrality of two subgraphs A and B are c_1 and c_2 respectively, and that of the reference graph is c_0. Then, $\Delta_b(c_0, c_1) < \Delta_b(c_0, c_2)$ means the second subgraph is more navigable than the first. **Propagativeness:** Propagativeness refers to a measure to capture how well information spreads through a subgraph S. Current flow betweenness centrality [12], based on Kirchoff's current laws, is designed to capture this concept. We combine this with the average neighbor degree of the nodes of S to measure the spreading propensity of S. Suppose the distribution of the current flow betweenness centrality of two subgraphs A and B is p_1 and p_2 respectively, and distribution of the reference graph is p_t. Also the distribution of the β_n, the average neighbor degree of the node n, for the subgraph A and B is γ_1 and γ_1 respectively, and the reference distribution is γ_t. If the condition $\Delta(p_t, p_1)+\Delta(\gamma_t, \gamma_1) < \Delta(p_t, p_2)+\Delta(\gamma_t, \gamma_2)$ holds, we can conclude that subgraph B is a faster propagating network than subgraph A. **Subgroups within a Candidate Subgraph:** The purpose of the last metric is to determine whether a candidate subgraph identified using the previous measures need to be further decomposed into smaller subgraphs. We use subgraph centrality [13] and coreness of nodes as our metrics. The subgraph centrality measures the number of subgraphs a vertex participates in, and the core number of a node is the largest value k of a k-core containing that node. So a subgraph for which the core number and subgraph centrality distributions are right-skewed compared to the background subgraph are (i) either split around high-coreness nodes, or (ii) reported to the user as a mixture of diverse topics. The node grouping, per-group subgraph generation and candidate subgraph identification process is presented in Algorithm 1. In the algorithm, function *cut2bin* extends the cut function, which compares the histograms of the two distributions whose domains (X-values) must overlap, and produces equi-width bins to ensure that two histograms (i.e., frequency distributions) have compatible bins.

IV. The Discovery Process

Algorithm 1: Graph Construction Algorithm

```
INPUT : Q_out Output of the query, L Graph construction rules, gv grouping
  variable, th_size is the minimum size of the subgraph;
Function gmetrics (Q_out, L, groupVar)
    G[]← ConstructGraph(Q_out, L);
    T ← [];
    for g ∈ G do
        t_α ← ComputeMetrics(g);
        T.push(t_alpha);
    end
    return T
end
Function ComputeMetrics (Graph g)
    m ← [];
    m.push(eigenVectorCentrality(g));
    ..... m.push(coreNumber(g));
    return m
end
Function CompareHistograms (List t_1, List x_2)
    s_g ← cut2bin(x_2, bin_edges);
    bin_edges ← getBinEdges(x_2);
    t_g ← cut2bin(t_1, bin_edges);
    β_js ← distance.jensenShannon(t_g, s_g);
    h_t ← histogram(t_g, s_g, bin_edges);
    return β_js, h_t, bin_edges;
end
```

Algorithm 2: Graph Discovery Algorithm

```
Input: Set of all subgraphs divergence σ
Output: Feature vectors v_1, v_2, v_3, List for re-partition recommendations l
ev : eigenvector centrality;
ec : edge current flow betweenness centrality;
nc : current flow betweenness centrality;
μ : core number;
z : average neighbor degree;
Function discover (σ)
    for any two set of divergence from σ_1 ans σ_2 do
        if σ_2(ev) > σ_1(ev) then
            v_1(σ_2) = v_1(σ_2) + 1;
            if σ_2(ec) > σ_1(ec) then
                v_2(σ_2) = v_2(σ_2) + 1;
                if (σ_2(nc) + σ_2(μ)) > (σ_1(ec) + σ_2(μ)) then
                    v_3(σ_2) = v_3(σ_2) + 1;
                end
                if (σ_2(sc) + σ_2(z)) > (σ_1(sc) + σ_2(z)) then
                    l(σ_2) = 1;
                end
            end
        end
    end
end
```

The discovery algorithm's input is the list of divergence values of two candidate sets computed against the same reference graph. It produces four lists at the end. Each of the first three lists contains one specific factor of interestingness of the subgraph. The most interesting subgraph should present in all three vectors. In the algorithm v_1, v_2 and v_3 are the three vectors to store the interestingness factors of the subgraphs, and l is the list for repartitioning. For two subgraphs, if one

	Group	Description
1	A	#ADOS Movement Related Group
2	B	American Political Group
3	C	General Black Related Issue
4	D	HIV, Drug etc. related
5	E	LGBT and Gay Issues
6	F	Random terms from Google top trends

TABLE II: List of Candidates with domain descriptions.

of them qualified for v_1 means, the subgraph contains high centrality than the other.

V. Experiments and Results

DATASET: The experimental dataset was gathered in the following manner. 1) We collected a set of tweets over a period of six months, such that the tweets use the the hashtag `#ADOS`, usually associated with Black American issues; 2) We adopt a snowball sampling strategy by which we identify the most very active users based on the number of tweets they author; 3) We collect all tweets from these users regardless of the topic content; 4) This process is performed for two more rounds. The size of the accumulated dataset is 9,780,590 tweets, and the number of unique users mentioned is 89,8850. A list of the keywords and the name of the collection is given in in Table II. The first column of the table is the group's name, and the second column represents the group's descriptions. In the candidate formation query, each group is represented by a set of keywords selected based on Google Trends such that these keywords cooccur with our seed keyword `ADOS`. *Node Grouping:* Initially, for each candidate, we grouped them using the popularity count of the tweet and the followers' count of the user. We have explored 10 different node groups, and the graph graphs are checked against our interestingness criteria. Furthermore, we empirically determine that attributes the tweet-popularity is a suitable the soft grouping variable is significant and practical to analyze because the followers count does not relate the content or the event directly. Hence we continue the experiment with tweet's popularity number as the grouping variable.

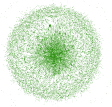
Experiments: We conducted experiments on the keyword categories shown in Table II. The first is directly related to the keywords used for data collection and the last one is randomly picked from Google trends with no relationship with the first. The remaining four have been selected as increasingly general issues found in Google Trends.

RESULTS: The primary observation about these graphs is that although they do have a perceptible nucleus-periphery structure, the width of the "peripheral ring" is thick and the space between the nucleus and the periphery is fairly crowded. The sample network in Figure 1b illustrates that some parts of the graph almost does not show any distinct peripheral boundary that establishes the strong edge formation probability between nucleus and non-nucleus nodes as well between a random pair of non-nucleus nodes. Given this backdrop, let us examine the subgraphs shown in Figure 1 – they are examples of positive and negative results from our algorithm.

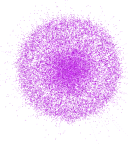
Subgraph 1. The subgraph shown in Figure 1d, characterized by a tight, strong core and a very scant periphery, is structurally interesting because it is significantly isolated from the rest of the network. Content analysis shows it is strongly focused on "Black" issues with extremely high interactions amongst users who have very little interest outside this narrow scope. The third tweet from Table I is the example of such tweets. In order to build a strong network community, they mention a small set of users numerous times, even without any content (first tweet of the same table). The network shows right-skewed eigenvector centrality distribution, high navigability (Group A in Figure 2a), but low propagativeness (Group A in Figure 2b). *Therefore Subgraph 1 is interesting.*
Subgraph 2. Figure 1e is an extensive network with a large and dense nucleus and a less dense but thicker periphery not very strongly connected to the nucleus. Like tweet 4 from Table I, people in the center wish to connect to strongly connected and focused network by mentioning other connected users and issues. As in tweet 5, people mention random unrelated users purposefully because it boosts their tweets' reach with loosely connected users, which creates a thick ring outside the kernel. From figures 2a, 2b, and 2c, we can see that it has very high navigability and propagativeness compared to the other groups. *Subgraph 2 is interesting as it discovers users who build bridges to promote message propagation*.
Subgraph 3. Figure 1f shows a network related to black issues (like healthcare) without specific focus on political issues. Hence, the network is not very intense (has a lighter nucleus), a peripheral density like Subgraph 2, and a diffuse space between then. Curiously, all our interestingness metrics score this subgraph highly. Upon closer inspection from Figures 2a, 2b, and 2c, we can see it has a spike on navigability, is well connected, and has a high propagativeness. The network also exhibits a high number of cores and subgroups *Hence we label this subgraph as interesting but not readlily interpretable*. So this network is considered for further partitioning. **Subgraphs 4 and 5.**. The networks shown in 1g and 1h are based on a deliberate choice of "general purpose" topics. Clearly, they have a lighter nucleus with a diffused ring, and fairly close to the random networks shown in the top 3 figures. This is confirmed by the low JS-divergence values for the navigability, propagativeness, and subgroup measures. *Hence we conclude these candidates are* ***not interesting*** *subgraphs*. **Subgraph 6.** Finally, Figure 1i produced from random set of keywords shows inconsistent results from our algorithm as the measures show no conclusive score on any metric that make it a proper interestingness candidate. *We therefore conclude that these subgraphs are* ***not interesting***.

VI. Conclusion

Our experiments show that the subgraphs that our algorithms report are indeed interesting. Our future work would involve making the algorithms more robust and devise a more elaborate evaluation methodology to validate the interestingness of the subgraphs recognized by our technique.

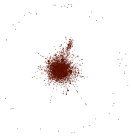


(a) Random Sample of the graph 1

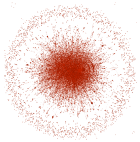


(b) Random Sample of the graph 2

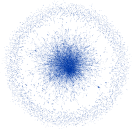
(c) Random Sample of the graph 3



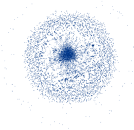
(d) #ADOS Movement Related Network filtered using #ADOS related Keywords.



(e) Political campaign-related network Based on the Presence of Political Personality.

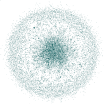


(f) Black Social Issues Network.



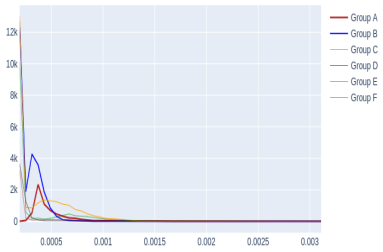
(g) HIV, Drug and PrEP relates Issues.

(h) LGBTQ Community Related Group.

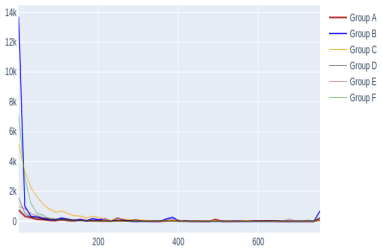


(i) Random Terms from Google Trend

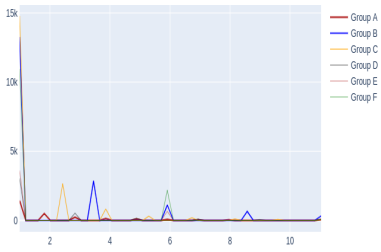
Fig. 1: User-Mention Network of Three different Data sets.



(a) Distributions of navigability.



(b) Distributions of propagativeness.



(c) Distributions of Subgroups in Candidates.

Fig. 2: Comparative Distributions of all Candidates.

Acknowledgment. Partially funded by NSF grants 1909875 and 1738411.

REFERENCES

[1] S. Sedhai and A. Sun, "An analysis of 14 million tweets on hashtag-oriented spamming," *Journal of the Association for Information Science and Technology*, vol. 68, no. 7, pp. 1638–1651, 2017.

[2] V. E. Lee, N. Ruan, R. Jin, and C. Aggarwal, "A survey of algorithms for dense subgraph discovery," in *Managing and Mining Graph Data*. Springer, 2010, pp. 303–336.

[3] X. Shan, C. Jia, L. Ding, X. Ding, and B. Song, "Dynamic top-k interesting subgraph query on large-scale labeled graphs," *Information*, vol. 10, no. 2, p. 61, 2019.

[4] F. Adriaens, J. Lijffijt, and T. De Bie, "Subjectively interesting connecting trees and forests," *Data Mining and Knowledge Discovery*, vol. 33, no. 4, pp. 1088–1124, 2019.

[5] P. Zhao, X. Li, D. Xin, and J. Han, "Graph cube: on warehousing and olap multidimensional networks," in *Proc. of the Int. Conf. on Management of Data (SIGMOD)*, 2011, pp. 853–864.

[6] M. van Leeuwen, T. De Bie, E. Spyropoulou, and C. Mesnage, "Subjective interestingness of subgraph patterns," *Machine Learning*, vol. 105, no. 1, pp. 41–75, 2016.

[7] X. Zheng and A. Gupta, "Social network of extreme tweeters: A case study," in *Proc. of the IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*, 2019, pp. 302–306.

[8] K. Das, S. Samanta, and M. Pal, "Study on centrality measures in social networks: a survey," *Social network analysis and mining*, vol. 8, no. 1, p. 13, 2018.

[9] P. Bonacich, "Some unique properties of eigenvector centrality," *Soc. Networks*, vol. 29, no. 4, pp. 555–564, Oct. 2007.

[10] B. Ruhnau, "Eigenvector-centrality—a node-centrality?" *Soc. Networks*, vol. 22, no. 4, pp. 357–365, 2000.

[11] X. Yan, Y. Wu, X. Li, C. Li, and Y. Hu, "Eigenvector perturbations of complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 408, pp. 106–118, Aug. 2014.

[12] U. Brandes and D. Fleischer, "Centrality measures based on current flow," in *Annual symposium on theoretical aspects of computer science*. Springer, 2005, pp. 533–544.

[13] E. Estrada and J. A. Rodriguez-Velazquez, "Subgraph centrality in complex networks," *Physical Review E*, vol. 71, no. 5, p. 056103, 2005.