Efficient Influential Researcher Selection

Jiongli Zhu*
jiz143@ucsd.edu

Shijie Sun*
shs001@ucsd.edu

Shuangyu Xiong*
s1xiong@ucsd.edu

Weiting Chen*
tic008@ucsd.edu

Abstract

With the growth of research communities in different research domains, developing an objective and reasonable way of selecting organizers as well as members to serve on program committees has gained more importance. In order to improve the attraction of the research conferences, we hope to select some influential researchers to become organizers or serve on the program committee of research conferences.

Based on this motivation, we propose to formalize the selection of influential researchers into a convex optimization problem based on real-world settings which aims at maximizing the total influence of the selected researchers with limitations on the number of selected researchers and the amount of price paid.

0 Task Assignment

- Data Preparation: Weiting Chen
- Algorithm Implementation: Shuangyu Xiong
- Experiments and Analysis: Shijie Sun, Shuangyu Xiong
- Literature Review and Paper Writing: Jiongli Zhu

1 Introduction

1.1 Motivation

In order to improve the attraction of the research conferences, we hope to select some influential researchers to become organizers or serve on the program committee of research conferences. We hope they can have both high academic achievement and strong connection in the research community, which can improve the quality of the paper review and benefit the advertising of the conference.

1.2 Related Works

Social network analysis often focus on macro-level models such as communities, diameter, clustering coefficient, small world effect, degree distributions, preferential attachment, etc; work in this area includes (Albert and Barabási, 2002), (Faloutsos et al., 1999), (Newman, 2003), (Strogatz, 2001). Recently, social influence study has started to attract more attention due to lots of important applications. Nevertheless, most of the works on this area present qualitative findings about social influences (Granovetter, 1973), (Nohria et al., 1992).

Several researches have been conducted to integrate publication data set and provided expertise researches search. For example, (Tang et al., 2008) extracted and integrated researchers' profile into a data set, and then provided expertise search. In the integration process, they utilized a probabilistic framework to address the name ambiguity problem. To provide expertise, they proposed three probabilistic topic models to simultaneously model the different types of information. There are also efforts made to retrieve influential nodes in social networks. For example, (Tang et al., 2009) proposed Topical Affinity Propagation to model the topic-level social influence on large network to generate topic-level social influence graphs for several topics. However, the above works focused on probabilistic models that are complicated and rely on the data distributions which are potentially biased.

On the contrary, (Elkin et al., 2013) tried to identify influential nodes by formulating and solving a convex optimization problem that is deterministic. Nevertheless, the problem was defined in a simple way and was not suitable to the context of influential research selection.

Similar to our work, (Bhushan et al.) tried to optimize the administration of COVID-19 vaccines by formulating it into a linear programming prob-

*Authors are alphabetically ordered by first names. All authors belong to the Department of Computer Science and Engineering.

lem. However, after its proposed simplification, the problem can be easily solved by an greedy algorithm which could be much more time efficient comparing to the solving the convex optimization algorithm. This greedy approach serves as *baseline-1* method in the experiments section.

1.3 Contributions

The contributions and novelty of our work can be concluded as follows:

- We formalize the selection of influential researchers into a convex optimization problem based on real-world settings in a novel way and derive its corresponding dual problem formulation and KKT conditions.
- We convert the original integer programming problem to a linear programming problem using relaxation, and empirically showed the tightness of this relaxation.
- We tested the proposed method on the Aminer dataset to evaluate: (1) the effectiveness of the selection of researchers, namely whether the selected researchers have incredible academic influence, (2) how k influences the runtime of the proposed method compared to baseline methods.

In the next section, we formulate the problem as a linear programming problem, set constraints based on real-world observations, and derive corresponding dual problem and KKT conditions. In section 3, we show details of the experiment implementation and compare the results of the proposed method as well as baseline approaches. In section 4, we further analyze the experimental results in terms of both effectiveness and efficiency. Finally, in section 5, we conclude and discuss the future directions of our work.

2 Methodology

Given an undirected unweighted co-authorship graph $\mathcal{G}$ with vector set $\mathbf{V}$ and edge set $\mathbf{E}$, where each node corresponds to an author. In the graph $v_i \in \mathbf{V}$ is defined as the total number of citation of the author i, and we also have $A \in \mathbb{Z}_2^{|V|\times|V|}$ which a binary $|V| \times |V|$ matrix denoting existence of edge between node pairs. More precisely, if an entry of A: $a_{i,j} = 1$, then there exists an edge between node i and j, and since $\mathcal{G}$ is an undirected graph, $a_{i,j} = a_{j,i}$.

2.1 Primal Problem Formulation

2.1.1 Optimization Problem

Our goal is to reasonably select influential researchers who are suitable candidates to serve on the program committee of research conferences. By influential, it not only means the researchers themselves have to possess excellent academic strength or achievement, which can increase the quality of the paper review., but they also need to have strong connections in the research community to benefit the advertising of the conference.

For simplicity, total citations can serve as the proxy of the academic ability, and connections can be evaluated by the number of coauthors. Since we value the connections between influential researchers, citations of coauthors are introduced as weights before counting. So the sum of citations of coauthors is chosen for measuring connections.

Based on the above intuition, we formalize the optimization problem as maximizing a weighted sum of two quantitative evaluation metrics:

$$\max_x (Av)^T x + \lambda \cdot v^T x$$

where $v \in \mathbb{R}^{|\mathbb{V}|}$. Besides, $x \in \mathbb{Z}_2^{|V|}$ is a binary column vector denoting which researchers are selected. By optimizing on the binary vector x, we can eventually obtain a reasonable selection maximizing our objective function. The first term of the objective function corresponds to the sum of citations of coauthors of selected authors, and the second term is simply the sum of citations of selected authors. Additionally, λ is some predefined non-negative balance constant.

2.1.2 Constraints

We can set constraints for the optimization problem based on some observations from real-world settings:

- Program Committee should have a capacity k.
- To avoid the case that inexperienced young researchers working in large groups are selected because of a large second term of the objective function, their respective citations should be no less than some lower bound l.
- Researchers with higher academic record should be hired with higher price. Therefore a limit of the amount of prices M should be set.

According to the above observations, we can formalize the corresponding constraints for the optimization problem:

$$\begin{aligned} &\mathbf{1}^T x \leq k \\ &p^T x \leq M \\ &e^T x = 0 \end{aligned}$$

where p is a $|V|$ dimensional vector in which $p_i = e^{\alpha v_i}$ as we assume the price grows exponentially with the number of citations and α is a constant scaling factor. Besides, $e \in \mathbb{Z}_2^{|V|}$ is an constant binary vector that satisfies:

$$e_i = \begin{cases} 0, & v_i \geq l \\ 1, & otherwise \end{cases}$$

This enforced the constraint that inexperienced authors with citations lower than l cannot be selected.

Besides, there is a hidden constraint for x since we need it to be binary, which makes our problem an integer programming problem. Given that the integer programming problem is harder than linear programming, we did a convex relaxation by converting this binary constraint to a continuous one:

$$\mathbf{0} \leq x \leq \mathbf{1}$$

This relaxation, according to (Elkin et al., 2013), is tight and the influencers (in our case are influential researchers) can be identified by solving the linear programming problem in an easier and efficient way.

We can further reorganize the primal problem with all its constraints in a classical linear programming form:

$$\begin{aligned} \max_x \quad & ((Av)^T + \lambda \cdot v^T)x \\ s.t. \quad & \begin{bmatrix} \mathbf{I} \\ \mathbf{1}^T \\ p^T \\ \mathbf{e}^T \\ -\mathbf{e}^T \end{bmatrix} x \leq \begin{bmatrix} \mathbf{1} \\ k \\ M \\ 0 \\ 0 \end{bmatrix} \\ & x \geq \mathbf{0} \end{aligned} \tag{1}$$

2.2 Dual Formulation

As discussed in (Seo, 2015; Gordon, 2016), the Lagrangian formulation of the linear programming problem (Eq. (1)) is:

$$L(x, y) = ((Av)^T + \lambda \cdot v^T)x + y^T(b - wx) \tag{2}$$

where

$$b = \begin{bmatrix} \mathbf{1} \\ k \\ M \\ 0 \\ 0 \end{bmatrix}$$

$$w = \begin{bmatrix} \mathbf{I} \\ \mathbf{1}^T \\ p^T \\ \mathbf{e}^T \\ -\mathbf{e}^T \end{bmatrix}$$

Therefore the corresponding Lagrangian dual problem is:

$$\begin{aligned} \max_y \quad & y^T b \\ s.t. \quad & w^T y \leq (Av + \lambda \cdot v) \\ & y \geq \mathbf{0} \end{aligned} \tag{3}$$

2.3 KKT conditions

KKT conditions can be used to test optimality. Based on the primal problem (Eq. (1)), we are able to derive the corresponding KKT conditions:

$$\begin{cases} wx^* \leq b \\ x^* \geq \mathbf{0} \\ L(x^*, y^*) = 0 \\ y^* \geq \mathbf{0} \\ y_i^* w_i = \mathbf{0} \quad for\ i = 1, \cdots, |V| + 3 \end{cases} \tag{4}$$

where x^* and y^* are optimal solutions returned by the convex optimization solver. y_i and w_i are the i^{th} element of y^* and the i^{th} row of w respectively.

3 Experiments

3.1 Experiment Settings

The algorithms in this paper were all programmed in Python(3.7.12), and were implemented on a x64 machine with 12G of memory, 114G of solid-state drive size, and both processors were Intel(R) Xeon(R) CPU @ 2.30 GHz (64 core), the operating system was Ubuntu 18.04.

The hyperparameters were set as follows:

- Maximum number of selected influential researchers k was set to 10.
- Balancing factor λ was set to 2.
- Upper bound of total price M was set to 18.
- Lower bound of individual citations l was set to 3,000.

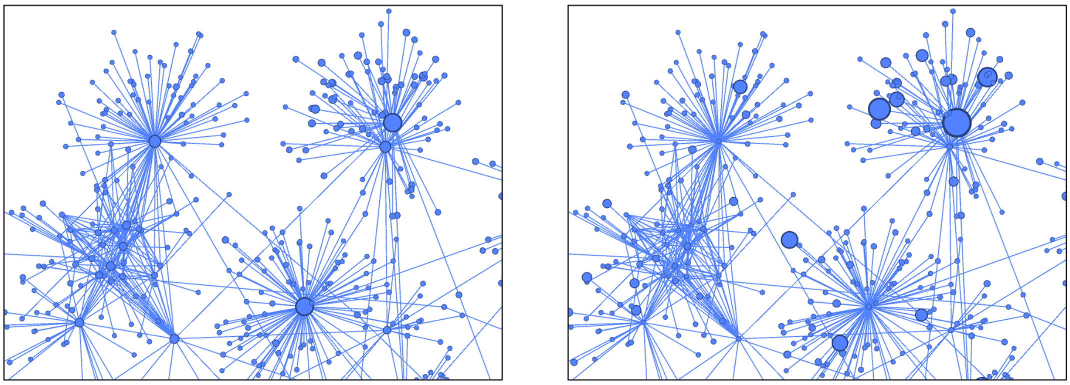

Figure 1: Visualization examples of the graph data. Node size grows with:
Left: influence, Right: price

- The constant scaling factor of citations α was set to 100,000.

3.2 Data Preprocessing

For the selection of datasets, the most direct idea was to crawl the academic websites such as google scholar, ACM, Digital Bibliography and Library Project, DBLP, etc. However, strong anti-crawler protection has made this infeasible. While in (Tang et al., 2008), the authors provided a way to obtain article data on DBLP. According to this method, we finally obtained as much article data on DBLP as possible.

After obtaining a large amount of data, our next goal is to preprocess the data, which was extremely time-consuming by using naive processing methods. If the dataset was reduced to a suitable size for experiments by random selection, authors from different domains would be selected which might result in a sparse graph with several connected components or even lots of individual nodes. Besides, it does not make sense to select influencers from authors with totally different research domains because it is almost impossible to have someone having connection with several irrelevant domains. Additionally, researchers from different domains are not likely to participate in the same conference.

To solve this problem, for each article, we stored a row for that article with all its authors, with the primary key set to a serial number that has no real meaning. Then for each author, we searched his or her coauthor based on his or her articles. After searching those coauthors, we constructed the coauthor table. The number of its entries in the article-author table is 5,239,262, which is extremely large, therefore we used SQLite (Hipp, 2020) which was already integrated with efficient large-scale data processing algorithms.

Metric	Value
#Nodes	1670
#Edges	2447
Average Node Degree	2.93
Median Node Degree	1
Diameter (longest shortest path)	5
Average Clustering Coefficient	0.05987
Total Citations	16492291
Average Citations	16012
Median Citations	6539

Table 1: Dataset Statistics

Finally, in terms of acquiring the citation values for each author, we relied on the open source python library, scholarly (Cholewiak et al., 2021), available on Github, to obtain the number of citations of scholars smoothly.

Since we were using a subset of the original large graph, the coauthors of some researchers might become incomplete. Therefore we added those missing neighbors (coauthors) of existing authors only for a more accurate computation, and those added authors (corresponding to 1-degree nodes) would not be involved in the influencer selection as their coauthors are incomplete.

3.3 Dataset Statistics

Table 1 gives a brief introduction of the dataset. Also, the distribution of researchers' prices is shown in Fig. 2, and we visualized the distribution of price and influence of researchers on part of the graph in Fig. 1, where the node sizes indicate the value of corresponding influence or price.

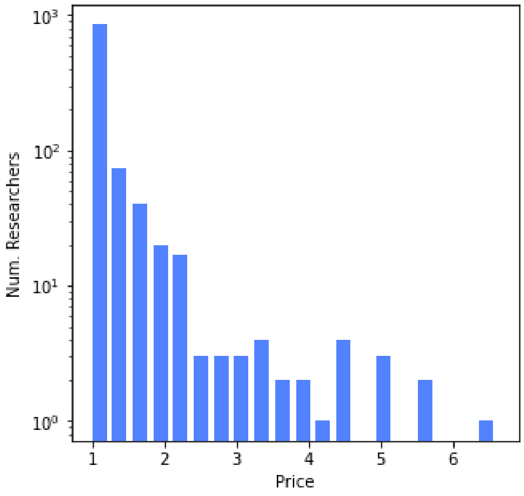

Figure 2: Value distribution of price vector p

3.4 Baseline Methods

To demonstrate the performance of the proposed optimization algorithm, we set up two heuristic methods as baselines.

Since it is impossible to test every feasible solution and find out the best node combination, a straightforward solution would be to find up to k top-ranked nodes till constraints no longer satisfied. That is, keep selecting nodes with highest s_i until k nodes are found or the total price reaches the limit M, where $s_i = (Av)_i + \lambda v_i$ is the contribution of single node can bring to the optimization goal if selected. We name this method as *baseline-1*.

For a stronger baseline, we can first narrow down the search space by heuristic rules and then figure out the best solution by brute-force searching. Specifically, we take c nodes ($c > k$) with highest s_i just like *baseline-1* and treat them as candidate nodes. Then we do exhaustive examinations over all the possible combinations between those candidates. After filtering out the combinations with all constraints satisfied (total price higher than M, no more than k nodes, etc.), we pick the one with the highest objective score. For simplicity, we take $c = 2k$. This is marked as *baseline-2*. Pseudocode of this approach is shown in Algorithm 1.

3.5 Experimental Results

We conducted experiments on the collected dataset using both our proposed linear programming

Algorithm 1: Baseline-2

```
Input: adjacent matrix $A$, citation vector $V$,
       price vector $P$, committee capacity
       $k$, balance constant $\lambda$, citation
       threshold $l$, budget $M$
Output: selected authors $selected\_ids$
$c = 2k$
$S = AV + \lambda V$
$cand\_ids = \text{set}()$
for $i = 1 \cdots |V|$ do
    if $V[i] \geq l$ then
        $cand\_ids.\text{append}(i)$
    end
end
$cand\_ids\_2$ =
  select_top_scored_ids($cand\_ids, S, c$)
$max\_obj = 0$
$selected\_ids = \text{set}()$
for $k' = 1 \cdots k$ do
    for
      $comb\_ids$ in all_comb($cand\_ids\_2, k'$)
      do
        if $\sum_{i \in comb\_ids} P[i] > M$ then
            continue
        end
        $obj = \sum_{i \in comb\_ids} S[i]$
        if $obj > max\_obj$ then
            $max\_obj = obj$
            $selected\_ids = comb\_ids$
        end
    end
end
return $selected\_ids$
```

method (described in Section 2) and two baseline methods (described in Section 3.4). The quantative comparison results which includes the value of objective function (weight sum that can somehow represent total influence of selected researchers), number of selected influential researchers and total price paid are shown in Tab. 2, and the visualization of selected nodes in the graph corresponding to three methods are shown in Figures 3 to 5.

4 Analysis

4.1 Performance Comparison

4.1.1 Effectiveness

As shown in Tab. 2, the greedy method *baseline-1* failed to balance between the goal of satisfying the price constraint and maximizing total academic

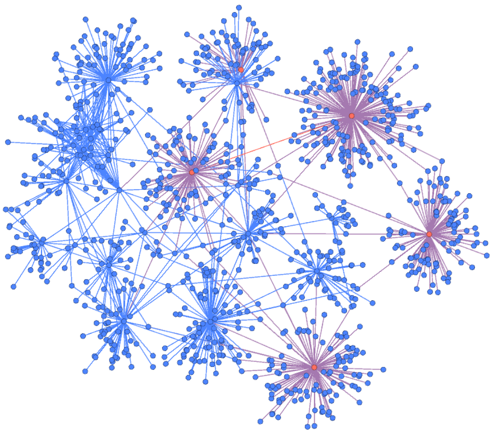

Figure 3: Visualization of baseline-1 selection

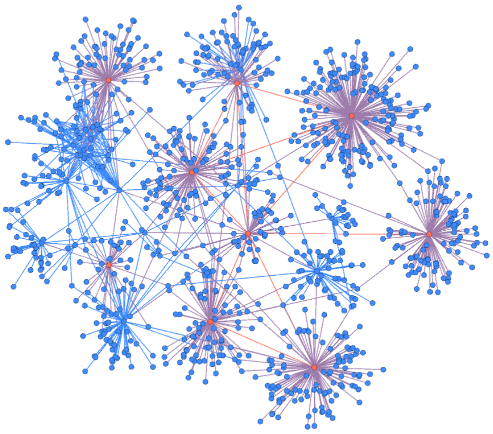

Figure 4: Visualization of baseline-2 selection

influence. This finally resulted in a lower final objective function value, and only 5 researchers were selected. Another baseline method set the number of selected researchers to k, but was just a stronger form of the previous greedy algorithm. Since the solution of *baseline-1* has been included in the search space of *baseline-2*, the result (value of the objective function) of *baseline-1* could be no better than *baseline-2*.

Since there was a limit M for total price, simply picking researchers with high individual influences would definitely exceed the limit, so we have to select a combination of researchers that have high individual influences (of course with high price) and researchers with relatively lower individual influences (but having much lower price). The aforementioned baseline algorithms were not sufficient to tackle this problem properly, resulting non-optimal selections.

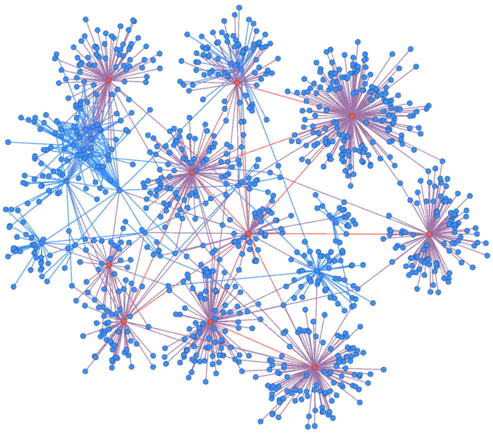

Figure 5: Visualization of the proposed selection

Method	Obj.	Num. Sel.	Price
baseline-1	1.04e7	5	17.06
baseline-2	1.44e7	9	16.63
proposed	**1.52e7**	10	17.85

Table 2: Quantitative Comparison Results

In contrast, as we would state in section 4.3, our proposed algorithm is able to maximize the objective function under the constraints and converge to the optimal solution. From the perspective of either quantitative (having higher objective function value in table 2) or visual comparison (establishing a broader connection in Figures 3 to 5), our proposed method outperformed other two baseline methods.

4.1.2 Scalability

As the limitation of selected researchers k may vary based on different needs in real-world scenarios, we conducted experiments to examine how does the runtime of algorithms changes with k. As shown in Fig. 6 and Tab. 3, the runtime of proposed method and *baseline-1* stay at a very low level, but the runtime of *baseline-2* grows at an extremely high rate with k. By analyzing Alg. 1, we can derive that the time complexity of *baseline-2* regarding k should be $O(2^{2k-1})$. In comparison, the runtime of the proposed method using convex optimization and the naive greedy approach *baseline-2* are sufficiently low and does not change much with k.

k	7	8	9	10
baseline-1	6.1ms	5.0ms	4.4ms	5.4ms
baseline-2	5.2s	19.9s	1min 9s	3min 2s
proposed	75.8ms	72.3ms	76.4ms	79.3ms

Table 3: Runtime Comparison

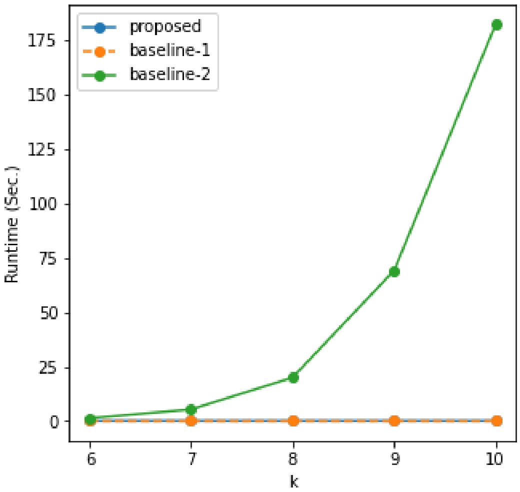

Figure 6: Runtime Comparison

4.2 Tightness of Relaxation

The histogram (Fig. 7) shows the distribution of the value in solution vector x, and this indicates that the values are either 0 or 1 in x, which eventually makes x a binary vector. Therefore the relaxation described in Section 2 is empirically proven to be tight enough in our context.

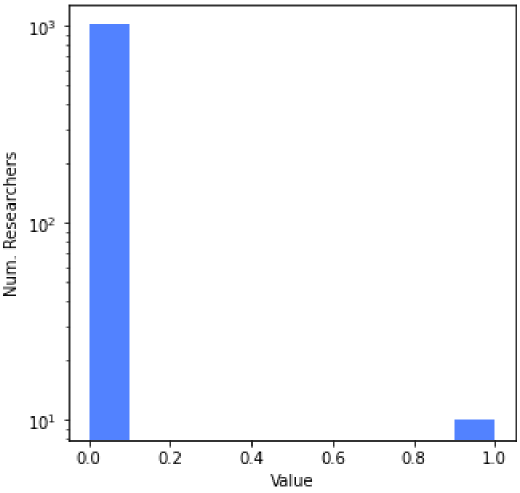

Figure 7: Value distribution of solution vector x

4.3 Optimality

Since our problem was a linear programming problem, there would always exist an optimal solution as long as the feasible region exists. By selecting only one researcher with price under limit we can get a valid solution (though might not be optimal), therefore the feasible set is obviously non-empty. We tested the optimality based on the KKT conditions described in Section 2.3, and all the conditions are satisfied using the returned x^* and y^*.

5 Conclusion

In this work, we developed an efficient algorithm for program committee selection. By introducing the citation and coauthor information, we formulate the problem as a convex optimization problem which maximizes academic influence of selected researchers. Through experiments under different settings, we show that the proposed convex optimization algorithm outperforms rule-based baseline. The results are also visualized and analyzed for a better understanding of the unseen pattern of research communities.

In the future, more factors such as h-index can be taken into consideration while formulating the problem. Besides, the influence of researchers can be further extended to two-, three- or even more hops neighborhoods in the graph, which can benefit obtaining more reasonable selections. Furthermore, this formulation of influence can be applied in other social network-related influencer selection problems with high efficiency and flexibility. The setting of constraints based on real-world observations might motivate the improvement of other existing works regarding academic social networks.

References

Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

Rohan Bhushan, Kunal Jain, Shivam Lakhotia, and Sothyrak Srey. Optimizing the administration of covid-19 vaccines.

Steven A. Cholewiak, Panos Ipeirotis, Victor Silva, and Arun Kannawadi. 2021. SCHOLARLY: Simple access to Google Scholar authors and citation using Python.

Lisa Elkin, Ting Kei Pong, and Stephen Vavasis. 2013. Convex relaxation for finding planted influential nodes in a social network. *arXiv preprint arXiv:1307.4047*.

Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. 1999. On power-law relationships of the internet topology. *ACM SIGCOMM computer communication review*, 29(4):251–262.

Geoff Gordon. 2016. Linear programming, lagrange multipliers, and duality.

Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.

Richard D Hipp. 2020. SQLite.

Mark EJ Newman. 2003. The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Nitin Nohria, Robert G Eccles, and Harvard Business Press. 1992. *Networks and organizations: Structure, form, and action*, volume 367. Harvard Business School Press Boston.

Hannah Seo. 2015. Lagrangean duality - optimization.

Steven H Strogatz. 2001. Exploring complex networks. *nature*, 410(6825):268–276.

Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. 2009. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 807–816, New York, NY, USA. Association for Computing Machinery.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: Extraction and mining of academic social networks. KDD '08, page 990–998, New York, NY, USA. Association for Computing Machinery.