Star Search:
Effective Subgroups in Collaborative Social Networks*

Ben Baumer†, George Rabanca‡, Amotz Bar-Noy§, and Prithwish Basu¶

Abstract—In an ever-increasing variety of contexts, people are working collaboratively to solve problems and accomplish tasks. Yet the characteristics of effectively functioning teams are not fully understood. We focus on the problem of identifying particularly effective teams in a large, complex, social-collaboration network. More specifically, given some task, and taking into account measures of both the effectiveness of an individual and the strength of the pairwise ties between individuals, can we identify the subgroup of people who are most likely to accomplish the said task? Our experimental work using DBLP data suggests that within any given subgroup, not all ties are of equal relevance. In fact, we show that focusing on the ties between one individual and the other group members—a star topology—will suffice. However, whereas the problem of finding the “best” subgroup is equivalent to MAX-CLIQUE and is thus hard to approximate, the problem of finding the best star is computationally tractable. We present experimental evidence justifying the interest in the star, as opposed to the clique, and discuss algorithmic and complexity concerns.

I. Introduction

The increasing complexity of the modern world brings with it many challenges. Whereas for much of human history, it was conceivable that an exceptionally well-educated person might be knowledgeable about virtually all scientific fields, and an expert in more than a handful, the depth of current scientific knowledge is such that most people can hope to truly master at most one sub-field of a particular discipline. That is, whereas Sir Isaac Newton was able to make landmark contributions to physics, mathematics, and astronomy, most scientists today specialize in a particular field within a particular discipline (e.g. data mining with computer science).

But while scientists have become increasingly specialized, the problems scientists face are increasingly complex, since the simpler problems have already been solved. Such complex problems (e.g. “how does an infectious disease propagate through the world’s population?”) touch upon work in multiple diverse fields (e.g. epidemiology, public health, microbiology, network science, statistics), and thus might attract the attention of researchers from multiple disparate disciplines. Indeed, it might be the case that such problems haven’t been solved precisely because a solution might require cutting-edge knowledge of multiple disparate fields, and no single person has that knowledge. Thus, it is by necessity that much of the work being done today—in science, but also in business, government, and industry—is done collaboratively in teams. Clearly, it would be simpler and more efficient to have a single person solve a single problem, but it is simply no longer feasible.

Today, while collaboration is ubiquitous, and it is clear that some teams work more effectively than others, the mechanisms that promote effective collaboration are not particularly well-understood. Such knowledge would be incredibly valuable in all walks of life, since the ability to create more effective and efficient teams would streamline production in any setting where teamwork is common. This would save two (one?) precious resources: time and money.

In this work, we construct a network-based mathematical model for collaborative work, use statistical learning techniques to motivate a focus on leader-based metrics for team evaluation, and discuss the algorithmic and complexity concerns inherent to this problem. As prescribed by much of the existing literature, incorporating all of the information about the pairwise connections between collaborators makes the problem of finding the most effective group computationally intractable. Instead, we demonstrate that the effectiveness of a team can be predicted nearly as well using only a vanishing fraction of these pairwise measurements. Moreover, by discarding most of the information, we make the problem computationally tractable.

A. Motivation: Our problem is motivated by the desire to understand how collaborative teams work towards solving problems. In particular, we are interested in: 1) within a particular team, how can we evaluate the cohesiveness of the team? 2) within a large network of potential collaborators, how can we identify teams that are likely to be highly effective?

We address these questions by considering a social network of potential collaborators, and a series of tasks that could be accomplished by teams comprised of these individuals. Each task has a quantifiable difficulty, or reward, and each collaborator has a quantifiable expertise or
skill. Furthermore, the pairwise strength of the connections between these collaborators in known. Thus, within this weighted social network, can we identify small subgroups of people who are likely to form effective teams? And what are the features of the social graph induced by the members of the team that importantly bear on the effectiveness of the Group A consists of five friends who work very well with each other, but have similar amounts of scientific expertise. Group B consists of an accomplished scholar, and four of his graduate students, none of whom know each other well, but all of whom are closely connected to the scholar. Is Group A more or less likely to accomplish a difficult task than Group B?

B. Related Work: Understanding what are the components that contribute to successful team formation has long been the focus of researchers in various disciplines. A great deal of effort has been put into empirical studies of team performance of various sizes in multiple areas. Levine and Moreland in [19] publish their findings on the performance of small teams and have concluded that their progress depends on both individual skill and group cohesiveness. Hauptman and Hirji [13] monitored individual behaviour and communication patterns in tens of projects from companies in various countries and industries and found a positive correlation between individual interdependence and team performance.

More recently, research in the operation research community has been directed at finding analytical models for good team selection in projects that require multiple individuals with various sets of skills. Zakarian and Kusiak [24] build a mathematical programming model in which team membership is prioritized based on costumer requirements or product characteristics: hard constraints the members of the team must satisfy, either individually or collectively. Baykasoglu et al. [4] develop a fuzzy model aiming to address the imprecise nature of the problem. Their model is optimizing for the inclusion in the team of a varying array of hard and soft skills since the actual performance of individuals is not generally known from the get-go. Citing previous research that provided arguments for linking successful teams with group communication, Chen and Lin [6] include in their model a personality profiling indicator in addition to teamwork capability, communication skills and domain specific knowledge of individual team members.

Although many of the models proposed before took into account the importance of group communication and team cohesiveness, these were generally modelled as a set of indicators of individual team members. It wasn’t until the work of Lappas, Liu and Terzi [18] that the pairwise collaboration, or compatibility, between any two members of the team has been considered. Since every pair of individuals is assigned a compatibility score, the pool of candidates for membership in the team can be seen as a social network. Besides raising some interesting algorithmic problems, their work also implicitly posed the question of what are the meaningful ways to aggregate the pairwise compatibility between different members of a team, to a team cohesiveness score.

Different aggregation functions lead to algorithmic problems that differ substantially in their computational complexity and the resulting team structure. While in [18] the authors argue a cohesive team has small diameter or minimum spanning tree in the social network subgraph induced by the team, in their experiments this assumption often leads to large teams in which many of the team members are selected just to minimize the shortest path distance between individuals who possess the required skills. A similar model is adopted by Datta, Majumder and Naidu [8] with the objective of minimizing the maximum edge in any Steiner tree. Gajewar and Sarma [9] adopts the subgraph density (the ratio of the number of edges to the number of vertices) as the metric by which to measure cohesiveness while Li and Shan [20] use the average (weighted) degree of each individual in the subgraph. However, all of those metrics still suffer from the same flaw: they don’t take into account the size of the team and often include individuals that don’t add any desired hard skills, but only make the team appear more cohesive.

In the model introduced by Kargar and An [16] the team quality is judged by the sum of the distances between each pair of individuals which in general leads to much smaller teams. They consider for the first time in this line of research teams with a leader, or what we refer to as the star topology, in which only the compatibility between the leader and the rest of the team is considered. They show that good teams with a leader are computationally easy to find, and also give algorithms for finding multiple teams (with or without leader) for independent projects. Rangapuram, Buhler and Hein also address the problem of large teams by presenting heuristics that take as input a bound on the team size to be specified, and a predetermined group of experts which are required to be part of the team. However, their solutions are only tested experimentally.

Other team formation literature has been concerned with online settings ([11]), where the goal is to allocate jobs fairly to individuals while still maintaining good team connectedness, minimizing the overall cost of the team ([17]) when each expert has an associated cost and pairwise compatibility reflects in the functioning cost of the team, or finding teams that can complete multiple projects, each one of them associated with a payoff ([10]).

The team formation model we propose leads to problems that are closely related to the graph packing literature. In the graph packing problem one is given an undirected, unweighted graph $G$ and a collection of undirected, unweighted graphs $\mathcal{G}$, and has the task of finding a collection $\mathcal{M}$ of disjoint subgraphs of $G$ such that every element of $\mathcal{M}$ is isomorphic to an element of $\mathcal{G}$. When $\mathcal{G}$ contains only a single edge graph, the graph packing problem is equivalent to the maximum matching problem. Hell and Kirkpatrick [14] showed that the problem is NP-complete for any collection $\mathcal{G}$ that is
not of the form \( \{K_{1,1}, K_{1,2}, \ldots, K_{1,T} \} \) where \( T \geq 1 \) is an integer and \( K_{1,t} \) is the star with \( t \) leaves. When the edges are weighted even this special case is NP-complete with the best known approximation algorithm developed by Babenko and Gusakov [3] guaranteeing a \( \frac{1}{2} - \frac{1}{e} \) approximation. Approximation algorithms for packing triangles in a graph have been presented in [7] while the packing of paths of sizes two and three has been studied in [11], [12], [22], [23] and [7].

C. Our Contribution: In this paper we make several contributions towards an understanding of how teams collaborate effectively. First, we propose a robust, versatile, and realistic mathematical model based on a social network architecture. Second, we use statistical evidence from real-world data to understand what features of the social network within a team are relevant to team effectiveness. Contrary to previous assumptions, we find that it is not necessary to consider all pairwise connections between collaborators in order to accurately predict the effectiveness of a group. Rather, the connections between the “leader” of the team and the other members are most important, the marginal predictive accuracy of the connections between other team members being small relative to the computational advantages of discarding most of the data. Third, we discuss how our method dramatically improves the computational complexity of the problem.

In Section II, we describe our mathematical model of team collaboration, and our statistical model for learning about how teams effectively collaborate. The results of our statistical modeling on the DBLP data set is presented in Section III, and our concluding remarks follow in Section IV.

II. Modeling

In this section, we present our mathematical model for collaboration on tasks. While our mathematical model is sufficiently general to model a variety of realistic scenarios, some of our language refers specifically to the setting of researchers writing papers collaboratively. This is simply a reflection of the fact that our primary data set is the DBLP, which contains information about authorship in computer science.

A. Mathematical Model: Let \( P \) be a set of workers, and let \( y \) be a set of tasks. Our goal is to understand how collaboration among the workers translates into success at completing tasks.

Among the workers we define a function \( h : P \to \mathbb{R}^+ \) that assigns an expertise value to each worker. In our DBLP example, we let \( h(p) \) be the h-index of the researcher \( p \). The \( h \)-index of a researcher is the largest integer \( H \) such that the researcher published at least \( H \) papers, each of which received at least \( H \) citations, and is generally believed to be a good indication of a researcher’s impact.

We measure two different types of collaboration among workers: pairwise collaboration between two workers, and team cohesiveness among three or more workers.

To measure pairwise collaboration between researchers we posit five different functions \( w_j : P \times P \to \mathbb{R}^+ \) and we compare results based on different choices for the collaboration function:

1) Binary. Have these two researchers collaborated on a paper:

\[
 w_1(p_1, p_2) = \begin{cases} 
 1 & \text{if } p_1, p_2 \text{ are co-authors} \\
 0 & \text{otherwise} 
\end{cases}
\]

2) H-index. Let \( A, B \) be the set of papers authored by \( p_1 \) and \( p_2 \), respectively. Then

\[
 w_2(p_1, p_2) = h\text{-index}(A \cap B).
\]

3) Citations. The total number of citations that all joint papers of the two researchers received. Let \( \text{cit}(A) \) denote the total number of citations papers in set \( A \) received. Then,

\[
 w_3(p_1, p_2) = \text{cit}(A \cap B).
\]

4) J-index. The Jaccard index of the paper sets of the two authors:

\[
 w_4(p_1, p_2) = \frac{|A \cap B|}{|A \cup B|}.
\]

5) J-citations. Jaccard index of the sets of citations received by the two authors:

\[
 w_5(p_1, p_2) = \frac{\text{cit}(A \cap B)}{\text{cit}(A \cup B)}.
\]

Thus, each pairwise collaboration function \( w_j \) defines a \textit{worker network} \( G_j = (P, 2^P, w_j) \), where edges of zero weight correspond to non-existent edges.

A team of workers is a subgraph of the worker network, and we assume any team has a \textit{collaboration topology}: an unweighted subgraph of the team graph in which an edge between two workers denotes the fact that the respective workers will have to collaborate. Therefore, we consider the following problem:

\textbf{Problem 1.} Given a weighted graph \( G = (V_G, E_G, w) \) and an unweighted graph \( T(V_T, E_T) \) find a one to one mapping \( g : V_T \to V_G \) which maximizes

\[
 f_T(g) = f(h(g(u_1)), h(g(u_2)), \ldots, h(g(u_k)), w_1(g(u_1), g(u_2)), \ldots, w_5(g(u_{k-1}, g(u_k))),
\]

where \( k \) is the number of vertices in \( T \).

In other words, we want to find a team of workers and assign them to the different nodes of the team graph (representing different job assignments), such that collaboration is maximized between all relevant pairs: the workers that are assigned to nodes connected by an edge in the graph \( T \). We call \( T \) the collaboration topology of the team, and \( f(g) \) the cohesive index of the worker assignment \( g \). Clearly,
the difficulty of computing an optimal worker assignment depends on the structure of graph $T$. We are interested in particular in comparing the advantages and disadvantages of creating teams with star versus clique collaboration topology. Observe that the star topology corresponds to a strictly hierarchical team structure with one leader and many independent workers, while the clique topology corresponds to a strictly flat, leaderless, team structure. While teams in real world organizations don’t usually have either a perfectly flat nor perfectly hierarchical structure, we find that, for clarity, is useful to compare and contrast these two extreme topologies. When the graph $T$ is a clique we will refer to problem 1 as the MAX-TEAM problem, and when the graph $T$ is a star we will refer to the problem as the MAX-STAR problem.

When the size of $T$ is $k$ and all the edges of $G$ are either 0 or 1, the MAX-TEAM problem is equivalent to deciding whether there exists a clique of size $k$ in $G$ (the MAX-CLIQUE problem) and is therefore NP-hard. However, the MAX-STAR problem can be solved in time polynomial in the size of the team and the size of $G$. A simple algorithm for this problem first finds for each worker $v$ the best star centered at $v$ by simply selecting the heaviest $k - 1$ edges of $G$ incident to $v$; then, it returns the best such star. Thus, while the problem statements are similar, their computational complexities are very different. The difference is obviously that in the first problem, all team subgroups—of which there are exponentially many—are equally important, whereas in the second case only the relation between the star center and the closest collaborators is important. The goal of our experimental section is to justify interest in MAX-STAR in the service of answering the question posed by MAX-TEAM.

More generally, we consider the question of how to meaningfully reduce the information embedded in any given subgraph of our social network into a single number. Our results suggest that a function which considers only a small fraction of the information in $G$ can be just as effective in addressing the real-world question of interest.

a) Clique Topology: One reasonable way [9], [20] to assess the collaborative strength of the team is to assume that all of the pairwise collaborations are relevant, and simply take the average of the edge weights. Since all of the edge weights (some of which may be 0) are included, we call this the 

clique topology.

$$f_{\text{clique}}(Q) = \frac{1}{\binom{|Q|}{2}} \sum_{(q_1, q_2) \in Q \times Q} w_{q}(q_1, q_2)$$

Note that using the clique topology to assess team strength implicitly asserts that all pairwise collaboration ties are meaningful. Although this may be reasonable, it has the downside of being computationally difficult. That is, to find the strongest team in a graph is equivalent to solving the MAX-CLIQUE problem, which is not only NP-hard, but also hard to approximate [15].

b) Star Topology: Conversely, we consider an alternative assessment of team strength based on the star topology. Here, we assert that only some of the ties between workers are meaningful. In particular, we assert that only the pairwise collaboration between a single team leader and the other members of the team are important. Critically, this dramatically reduces the complexity of the computational problem, since finding the maximum-weighted star can be solved in polynomial time.

Let $q^* \in Q$ be the team leader. Here, $q^*$ is defined by $q^* = \max_{q \in Q} h(q)$. Then

$$f_{\text{star}}(Q) = \frac{1}{deg_{G^*(Q)}(q^*)} \sum_{q \neq q^* \in Q} w_{j}(q^*, q)$$

B. Statistical Model: While the preceding discussion presents computational motivation for focusing on stars as opposed to cliques, this motivation is pointless if collaboration cannot be meaningfully characterized on the star topology. Thus, we perform statistical analysis on the DBLP data set that justifies our focus on stars.

Generally, we consider the problem of learning a function $f: G^2 \rightarrow \mathbb{R}^+$ that models collaboration accurately.

We consider this problem in the context of machine learning. That is, given a vector of outcomes $y$, and a matrix of attributes $X$, what function $f$ minimizes the (squared) difference between $f(X)$ and $y$? The difficulty is that in our problem $X$ is not a matrix—but rather it is a subgraph with weighted nodes and weighted edges. What function most accurately aggregates this information into a prediction?

In our analysis, $y$ is the number of citations earned by papers listed in DBLP. Thus, we must learn a function $f$, taking as its input a weighted subgraph of authors and their pairwise connections, that predicts the number of citations that a paper written by that group of authors would produce.

The function $f_{\text{star}}$ presented above is one such function, but most likely it is not the best one. We tested a variety of candidate functions with various properties.

The performance of functions based on the star topology leads us to consider the possibility that only a few of those “spokes” are in fact meaningful. With $q^*$ defined as above, let $w_1, w_2, \ldots, w_k$ be the edge weights between $q^*$ and every other member of the team, sorted from largest to smallest. Then we consider the regression model

$$f_{\text{star,reg}}(Q) = \beta_0 + \beta_1 \cdot w_1 + \cdots + \beta_k \cdot w_k + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ for some constant variance $\sigma^2$. We use training data from the DBLP to estimate the values of the $\beta_i$'s.

1) Validation: For any collaboration function $f$, we can measure the root mean squared error (RMSE) of the
predictions made by the function. That is,
\[
RMSE_f = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (f(Q_i) - y_i)^2}
\]
where \(Q_i \subseteq P\) is the set of authors of the \(i^{th}\) paper, and \(y_i\) is the number of citations earned by the \(i^{th}\) paper.

Functions that produce smaller \(RMSE\) are more accurate, and are thus preferred. Our analysis demonstrates that collaboration functions based on the star topology are competitive with those based on the clique topology.

**Problem 2. Collaboration:** Given a weighted network \(G = (V, E, w)\), a vector of outcomes \(y\), and a map \(M : 2^G \rightarrow y\) that maps subgraphs of \(G\) to elements of \(y\), find a function \(f : 2^G \rightarrow \mathbb{R}^+\) that minimizes \(\|f(Q) - y\|_2^2\) (e.g. \(RMSE_f\)) over all subsets \(Q \subseteq V\).

### III. Experimental Results

In this section we describe our efforts to learn about the collaboration function that translates attributes of researchers in the DBLP into predictions about how many citations their papers will earn. The DBLP contains data about computer science authorship, and we obtained data about the number of citations earned by each paper from [2].

#### A. Small Teams:

In the first experiment we evaluated how well different combinations of pairwise collaboration functions and team collaboration topology assumptions predict the strength of a team. The data set we used for this experiment consists of 24,020 papers with at most five authors published in eight journals and twenty conferences in theoretical computer science between 1954 and 2002. This data was then further divided into two subsets: training data consisting of 21,006 papers published from 1954-2000, and testing data consisting of the remaining 3,014 papers published in 2001-2002.

We evaluate each of the combinations of the pairwise collaboration function with a team collaboration aggregation function compared to the observed strength of a team: the number of citations the newly published paper received. We use the Pearson product-moment correlation coefficient (PCC) to assess how good the author collaboration and aggregation functions are. The PCC of two random variables is a number between \(-1\) and \(1\) describing how the variables are correlated. In our setting, a PCC of 0 would suggest that team cohesiveness does not have any predictive power in identifying successful teams. The PCC of two random variables is defined as the covariance of the two variables divided by the product of their standard deviations:

\[
\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}
\]

Table I shows how the prediction of the team strength based on previous collaboration of authors is related to the number of citations of respective papers. Each row shows the results for a different function for computing the pairwise author collaboration, as described in section II. The columns show the results when different team collaboration topologies are considered.

In each cell we list the Pearson correlation coefficients between the number of citations of a paper and the number of citations their papers will earn. For example, when the pairwise collaboration between two authors is calculated as the total number of citations the authors received from common papers, and the team collaboration topology is assumed to be a clique, then the team cohesiveness score had a .194 Pearson correlation with the number of citations the paper received. It is interesting to notice that the best correlation is achieved in the setting that considers the least amount of information: the star topology with the binary pairwise collaboration function. The binary collaboration function is the simplest, in that it only records whether two authors collaborated previously or not, without taking into account the performance of the teams in which the authors collaborated previously. Besides, the correlation coefficient is always better when the star topology is assumed than when the collaboration between all pairs of authors are given equal importance.

It is surprising as well to observe that some author collaboration functions don’t perform as well as one would expect. The Jaccard index based collaboration functions do not show a relevant correlation to the number of citations in any of the team collaboration topologies. This collaboration function has been used in the experimental evaluation sections of [9] and [18]. To get a perspective on the relevance of the correlation coefficients in Table I note that a (positive or negative) Pearson correlation coefficient of .037 can occur by coincidence in unrelated distributions with a probability of about 10\%, while a coefficient of .15 can occur in unrelated distributions with a probability of \(10^{-5}\).

We also analyzed the the correlation between the number of citations and the sum of the author expertise or \(h\)-indexes to obtain a better correlation: .323. This confirms the common sense expectation that individual expertise is most crucial to the success of a team, but, the small difference between this correlation coefficient and the best correlation coefficients in table I suggests that choosing team members that collaborate well is almost as important as choosing team members with high expertise.

<table>
<thead>
<tr>
<th></th>
<th>star</th>
<th>clique</th>
</tr>
</thead>
<tbody>
<tr>
<td>h-index</td>
<td>.150</td>
<td>.130</td>
</tr>
<tr>
<td>citations</td>
<td>.195</td>
<td>.194</td>
</tr>
<tr>
<td>j-index</td>
<td>.007</td>
<td>.013</td>
</tr>
<tr>
<td>j-citations</td>
<td>.034</td>
<td>.037</td>
</tr>
</tbody>
</table>

**TABLE I.** PPC BETWEEN TEAM COHESIVENESS AND NUMBER OF CITATIONS WITH DIFFERENT PAIRWISE COLLABORATION METRICS AND AGGREGATION FUNCTIONS.
B. Large teams: This data set consists of 622,094 papers published in a wide array of venues between 1990 and 2003. From this data we were able to construct a graph \( G \) with 43,453 authors, having between them 181,812 pairwise connections, and authoring 240,621 papers.

Given the larger size of the latter data set, we focus our presentation here on those results. However, the results from the smaller data set were consistent.

As noted in Section II, we take as input a large social network \( G \), where the nodes are weighted by the h-index of each author on all papers published before 2001. Similarly, the pairwise connections between authors were computed as the h-index of the shared publications of those authors on all papers published before 2001. Thus, the background data embedded in \( G \) consists entirely of publications before 2001.

The results of testing six regression-based prediction models are shown in Table III and Table IV. Each model is evaluated in terms of its ability to estimate the logarithm of the number of citations of each paper (+1). That is, the response variable is \( \ln(citations + 1) \) and thus the units of the RMSE figures are in log-citations. The models are as follows:

- **mean**: assign every paper the mean number of citations. This serves as our baseline.
- **year**: since papers only accrue more citations over time, use the year in which the paper was published to predict its citations. Here, both a linear and quadratic term for year are allowed.
- **numAuthors**: use the number of authors on the paper to predict citations. It is assumed that there is no correlation between citations and number of authors, but this serves as a useful verification of a random noise model.
- **density**: \( \text{density}(G) = \frac{1}{|E(G)|} \sum_{e \in E(G)} w(e) \). This captures the idea from the clique topology that all edges are important, and that teams with stronger pairwise connections will perform better.
- **star**: \( \text{star}(G) = \max_{v \in V(G)} \frac{\sum_{w \in \text{deg}(v)} w^2}{{\text{deg}}(v)+1} \). This captures the idea from the star topology that only the mostly strongly-connected member of the team is important.
- **topTwo**: use the largest two h-indices and the largest two edge weights in the subgraph to predict the number of citations. While not a star topology, this captures the notion that only the two best researchers and two strongest ties are important. The magnitude of the other weights in the subgraph are less important.

An example illustrating these collaboration functions is shown in Figure 1. Here, we consider the paper “A community authorization service for group collaboration”, published in 2002 by Laura Pearlman, Von Welch, Ian T. Foster, Carl Kesselman, and Steven Tuecke [21]. The h-indices and joint h-indices of the five co-authors prior to 2001 are indicated numerically in the figure. This paper has received 504 citations.

![Fig. 1. An example team graph.](image)

<table>
<thead>
<tr>
<th>( f )</th>
<th>score</th>
<th>predicted citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>17.5</td>
<td>17.5</td>
</tr>
<tr>
<td>year 2001</td>
<td>15.4</td>
<td>17.6</td>
</tr>
<tr>
<td>numAuthors</td>
<td>5</td>
<td>17.4</td>
</tr>
<tr>
<td>density</td>
<td>7.3</td>
<td>16.6</td>
</tr>
<tr>
<td>star</td>
<td>5.7</td>
<td>16.6</td>
</tr>
<tr>
<td>topTwo</td>
<td></td>
<td>36.5</td>
</tr>
</tbody>
</table>

Table II shows the predictions of the six models above for the paper considered in Figure 1. The mean citations predicted by the baseline model is 17.5 (for all papers). In this case, only the topTwo model performs significantly better than that. While the scale of the scores is irrelevant, and the details of how these scores translate into predicted citations are omitted, what should be clear is how the topology of this group is reflected in the different scores. In this example, three people with high h-indices have very strong connections, while two people appear to new to the group, and perhaps computer science research in general. Although in this example, the density metric performs slightly better than the star, it is clear how the strength of a few people can drive the success of the group.

Table III compares the performance of each of these function on the training data, which consisted of 15,845 papers which had at least three authors. As expected, year and numAuthors are largely indistinguishable from the baseline model that simply uses the mean. Moreover, the model based on density is not much better. However, predictions based on the star metrics offer a considerable improvement, and focusing on the topTwo is even better.

In Table IV, we compare the performance of these same functions on a hold-out set of 13,753 papers on which they were not trained. The conclusions are largely the same. While the density metric performs better than the baseline model against this sample, it is still surpassed by the star and topTwo models.

In Figures 2 and 3, we show the bivariate relationship between the actual number of citations earned by each paper, and the team cohesiveness score with the star metric.
TABLE III. RESULTS FROM IN-SAMPLE TESTING OF VARIOUS COLLABORATION FUNCTIONS ON THE DBLP. EACH FUNCTION WAS TRAINED ON 15,845 PAPERS AND EVALUATED ON THOSE SAME PAPERS.

<table>
<thead>
<tr>
<th>Function</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>1.705</td>
</tr>
<tr>
<td>year</td>
<td>1.703</td>
</tr>
<tr>
<td>numAuthors</td>
<td>1.705</td>
</tr>
<tr>
<td>density</td>
<td>1.711</td>
</tr>
<tr>
<td>star</td>
<td>1.669</td>
</tr>
<tr>
<td>topTwo</td>
<td>1.540</td>
</tr>
</tbody>
</table>

TABLE IV. RESULTS FROM OUT-OF-SAMPLE TESTING OF VARIOUS COLLABORATION FUNCTIONS ON THE DBLP. EACH FUNCTION WAS TRAINED ON 15,845 AND EVALUATED ON 13,753 DIFFERENT PAPERS.

<table>
<thead>
<tr>
<th>Function</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>1.706</td>
</tr>
<tr>
<td>year</td>
<td>1.690</td>
</tr>
<tr>
<td>numAuthors</td>
<td>1.708</td>
</tr>
<tr>
<td>density</td>
<td>1.710</td>
</tr>
<tr>
<td>star</td>
<td>1.695</td>
</tr>
<tr>
<td>topTwo</td>
<td>1.582</td>
</tr>
</tbody>
</table>

and density metric respectively.

IV. Conclusion

In this section we present the limitations of our work, explore avenues for future study, and summarize our findings.

A. Limitations & Future Work: We have explored only a small portion of the infinite space of potential collaboration functions. Our motivation in this effort was to find collaboration functions that led to computationally tractable algorithms that performed as well or better than exponential time algorithms necessary for MAX-TEAM. In that effort we have been successful, but it is certain that these collaboration are not optimal. A worthwhile goal for future study would be to explore the infinite space of potential collaboration functions to find an optimal one.

Here, we have considered only the DBLP data set, which is specific to computer scientists writing academic papers. Thus, our conclusions about the usefulness of the star topology relative to the clique topology apply narrowly to this setting. While we suspect that finding likely extends to other application domains, we have no evidence other than intuition to support this claim.

One challenging aspect of this data set was the heavily right-skewed distributions for h-index and number of citations. This was especially troublesome for the response variable of number of citations. While our remedy of taking the logarithm was an improvement, a more careful study might explore more sophisticated statistical transformations. In particular, both quantities appear to follow zero-inflated negative binomial distributions (see Figure 4). It seems likely that successfully transforming these variables could result in more accurate predictions.
**B. Summary:** The question of how people collaborate most effectively is both important and not fully understood. We used the DBLP database of computer science authorship to address this question. Whereas previous research focused on the strength of ties among all members of a team, we focused on the strength of ties among only some of the members of a team. In particular, we used the DBLP to provide experimental justification for a focus on the star topology over the clique topology. This experimental evidence suggests that metrics based on the star topology are competitive with those based on the clique, but are computationally tractable (in contrast to our prior work on finding cliques with a strong weakest link [5]). This makes a practical case for using the star topology when mining for effective subgroups in large social networks.

**References**


