Analysis on Communication Cost and Team Performance in Team Formation Problem

Weijun Chen, Jing Yang and Yang Yu*

School of Data and Computer Science, Sun Yat-sen University,
Guangzhou, P.R.China
chenwj96@mail2.sysu.edu.cn, yangj357@mail2.sysu.edu.cn,
yuy@mail.sysu.edu.cn

Abstract. Team formation problem refers to finding a set of skillful individuals to accomplish given tasks as a team. A growing interest of recent researches on team formation is to concern the collaboration factor, following the general idea that effective communication among team members may contribute to better team performance. Previous studies have introduced a variety of ways to model the collaboration factor applying the concept communication cost, yet few have investigated the effectiveness of the proposed metrics. In this paper, an empirical study is conducted to evaluate the effectiveness of existing communication cost metrics in terms of the influence on team performance. We select real data from IMDb as an example for the study and apply statistical analysis. Based on the result of evaluation, we further propose modification for the communication cost metrics and demonstrate the feasibility. This empirical study is expected to provide suggestion and inspiration for both researchers and practitioners, while design or select models to solve team formation problem in different application scenarios.

Keywords: team formation, collaboration, communication cost, social network metrics.

1 Introduction

Team formation problem refers to finding a team of actors (i.e. human resources) to accomplish a project with certain constraints such as skill qualification, business roles, etc. The problem of team formation origins from the Operations Research. One may find various real-life scenarios for the application of team formation, such as software development, task distribution in workflow management, partner selection in virtual enterprises, etc.

There is a growing interest on the study of the collaboration among individuals in group work and its potential impact on team performance [1,2]. Lappas et al. addresses the problem of team formation concerning the collaboration of team members [3], which states that effective cooperation contributes to better team performance and should therefore be considered besides the requirements for skills. The concept of

* Corresponding author.

communication cost is used to measure the effectiveness of cooperation within a team, and may be calculated using techniques from social network analysis, based on the social network model that captures the interactions of the individuals. Later researches in the same stream have proposed different network-based metrics for measuring communication cost, depending on the application scenarios of group work. Despite the well-recognized existence of social network in real-life, these scenarios may vary in terms of the interaction patterns. For example, in a team that requires every team member cooperating smoothly with one another during group activity, such as surgical teams in health care [4], the Sum of Distance metric [5] may be used for modelling the communication; while for a leader-centric team on the other hand, the interactions between ordinary team members and team leader may be more important, thus Leader Distance metric [5,6] is more appropriate for capturing the pattern. The selection of metrics for measuring communication cost of members within teams would require domain knowledge in practical use.

However, while most of the researches focus on modeling the diversity of application scenarios, few have evaluated the effectiveness of these metrics. Although the concept of higher communication cost bringing adverse effect is well-accepted, it remains unclear how much communication cost would influence the performance of a team. On the other hand, most of the existing researches recognize different communication structures of teams as abstract graph patterns, while practitioners face more complex structures and may use more specific suggestion during the modeling phase. There is need for methods of evaluating the effectiveness of these communication cost metrics.

In this paper, we seek to achieve better understanding on the influence of communication cost in team formation problem through an empirical study. We select the IMDb movie data [7] commonly adopted in previous researches as one real example for case study. Statistical analysis is used to discover the potential correlation between four typical communication cost metrics and the performance of teams. By leveraging the results of analysis, we also provide suggestion on designing more effective metrics for measuring team communication cost. The presented research is expected to serve as a supplement to the current study on team formation problem concerning the collaboration factor, and may contributes to team formation in practical use.

The rest of this paper is organized as follows. Section 2 introduces related work on team formation problem, and from which four typical types of communication cost metrics are selected for later analysis. Section 3 reports the results of the empirical study that aims to investigate the correlation between communication cost and team performance. The suggested improvements on the metrics are demonstrated in Section 4. Section 5 givens a brief conclusion of the presented study.

2 Related Work and Preliminaries

Cooperative behaviors in teamwork affect service quality and performance [1], [8]. Related researches on team formation have taken into account communication cost among team members [3], [5,6], [9,10,11,12,13]. Reference [3] first considered communication as one important factor affecting the effectiveness of a team, and

formulated team formation problem based on communication cost functions on diameter distance and minimum spanning tree distance. [5,6] further extended previous researches and proposed team formation algorithms based on communication cost functions on the sum of distances and leader distance. [9] considered both communication cost and personnel cost of team members. [11] proposed an algorithm to optimize both individual expertise and handover relations to form a required team of high performance. [12,13] proposed algorithms to ensure the balanced workload of team members. [10] introduced the concept of capacity of actors, adding the constraint of capacity as the maximum workload allowed. [14,15,16] argued that team members would cooperate better if they have closer relations and tried mining social relations from historic data. [17] summarized recent team formation researches and developed a platform called Unified System for Team Formation to show the performance of different team formation algorithms.

The collaboration among potential members in team formation can be modeled as a social network graph $G(V, E)$, where V represents the set of all individuals and E is the set of direct association between two different individuals v_i and v_j, denoted as $e(v_i, v_j)$. Assign weights to the associated edges with social distance value, then the communication cost of v_i and v_j in G is therefore determined by the shortest path distance between v_i and v_j, denoted as $sp(v_i, v_j)$. In the following chapter, we study to investigate and compare the following metrics [3], [5,6] commonly adopted for measuring communication cost of a team X, which is a subset of V and form subgraph G':

- $Cc - R(X)$: The communication cost of a team X on diameter distance is defined as the diameter of G', which is the longest shortest path length between any pair of members in X.
- $Cc - MST(X)$: The communication cost of a team X on minimal spanning tree (MST) distance is defined as the cost of the minimum spanning tree on G'.
- $Cc - SD(X)$: The communication cost of a team X on the sum of distances is defined as the sum of all $sp(v_i, v_j)$ for all pairs of members (v_i, v_j) in X.
- $Cc - LD(X)$: The communication cost of a team X on leader distance is defined as the sum of all $sp(v_i, v_L)$ in X, where v_L is the team leader and v_i is any other member.

3 Correlational Analysis

The movie data from Internet Movie Database (IMDb) are commonly adopted by previous researches for evaluating the feasibility of the proposed algorithms [5], [9], [12,13]. We choose it for case study and apply statistical analysis to discover about the potential correlation between communication cost and the performance of the team.

3.1 The IMDb Dataset

IMDb dataset provides massive records about movie information. Each of the record contains information including y*ear, title, directors, actors* and *ratings* of a movie.

Specifically, the *ratings* of a movie are given by the users of the IMDb website, ranged from 0 to 10. Movie data from 2010 to 2016 are selected, containing a total of 132936 movies.

Table 1. Examples of the experimental dataset.

Movie ID	Year	Ratings	Team members
4027597	2016	7	[578448, 208647, 44352, 82602, 546937, 273875, 781837, 624307, 338718]
4028026	2016	6.5	[702169, 132988, 826693, 306446, 4444, 89452]
4028061	2016	7.8	[276204, 415364, 29396, 51812, 151432, 799030, 224061, 20734, 331220, 465837, 126350, 3434, 622141]
4028295	2016	4.7	[183758, 507130, 219560, 128275, 298734]
4027597	2016	7	[578448, 208647, 44352, 82602, 546937, 273875, 781837, 624307, 338718]

We follow the similar experimental setup [5], [9], and take directors and actors as individuals in the group work (of producing movies), and pick only the "skillful" individuals who have participated in at least 5 movies. Given a team, the ratings of the produced movies may serve as the result of a relative objective evaluation of the team's performance. The extracted dataset for analysis includes 89800 skillful individuals, who combined and produced 79669 movies. Examples of the dataset are shown in Table 1, where numerical ids are used to replace titles and names of persons.

We use the history of cooperation in producing movies (i.e. simultaneous appearance in the cast) to construct the collaboration network $G(V, E)$ of the directors and actors. Accordingly, the social distance value $e(v_i, v_j)$ could be determined by utilizing the movie ratings:

$$e(v_i, v_j) = 10 - avg\_ratings(v_i, v_j) \ , \tag{1}$$

where $avg\_ratings(v_i, v_j)$ represents the average of ratings of all movies in which v_i and v_j cooperate. The definition of $e(v_i, v_j)$ can be interpreted as: with past experiences of cooperating as team members and performing well, two individuals are expected to have shorter distance and thus communicate better.

3.2 Analysis of Correlation

To perform the analysis, split the dataset into two parts and pick the earlier one (ranged from year 2000 to 2009) for constructing the collaboration network of individuals, while the other (2010 – 2016) is left for observation. The four selected communication cost metrics are calculated respectively for each of the teams in the observation data. For teams in the observation data, we calculate the communication cost measured by each of the four selected metrics. Depending the definition of these graph-based metrics, it requires that the subgraph network G' of a team being connected. Thus we omit the records of unqualified teams during calculation and keep only those on which all four metrics could be applied.

The obtained results along with the corresponding movie ratings are used for correlational study. Pearson correlation coefficient (PCC) is adopted for analysis, which is a measure of the linear correlation between two variables. It also applies to a sample: Given two sets of sampling data of size n, $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$, the sample PCC denoted by r can be calculated as follows.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} , \quad (2)$$

where x_i, y_i are the single samples while $\bar{x}$, $\bar{y}$ being the sample means respectively. The value of r lies between +1 and -1, where +1 indicates total positive linear correlation between the two sets of data, and -1 for total negative correlation, and 0 for no linear correlation.

Table 2 demonstrates the results of correlational analysis using Pearson correlation coefficient. For all four metrics, negative PCC values were obtained, which imply negative correlation between the measured communication cost and the performance of the teams in observation data. In terms of the size of correlation, we found the PCC values relatively low. This is expected since the movie ratings are adopted to evaluate team performance, while there may be various other complicating factors other than the communication cost among team members. Nevertheless, the PCC value still vary to certain extent for these four metrics. By comparison the absolute value of PCC of $Cc - R$ is greater than that of $Cc - SD$, indicating that communication cost on *diameter distance* has stronger negative linear correlation with the team performance in this specific context. The result of comparison suggests that we may choose $Cc - R$ over $Cc - SD$ while modeling team formation problem in this case, since it is more confident that communication cost measured by $Cc - R$ would follow the tendency "as communication cost increases, the performance of team decreases".

The above analysis could be applied as a part of the solution to team formation problem in different contexts. After modeling the collaboration, one may leverage the historical results of group work by the focused individuals, and then determine whether the current modeling approach needs improvement or not. The results of group work are assessed through the use of Key Performance Indicators (KPI), which is a common practice in many fields and applications. Analysis should be done from case to case, and consider the specific characteristics of cases that may influence the analysis process.

In the following chapter, we continue the study on the movie production case, and introduce the modification on the existing metrics that may contribute to more effective measuring of communication cost.

Table 2. PCC between four communication cost metrics and team performance for observation data with 3623 records.

Communication cost metric	Pearson correlation coefficient
$Cc - R$	-0.34293
$Cc - MST$	-0.27914
$Cc - SD$	-0.10675
$Cc - LD$	-0.24793

4 Design of More Effective Metrics

4.1 Modifying the Original Definitions

Recall that at the end of chapter 2, the communication cost between any two individuals v_i and v_j in G is defined as the shortest path distance $sp(v_i, v_j)$. We further consider the transitive characteristic of social relationships, and make the following assumption on the communication pattern among individuals: if v_i and v_j communicate and collaborate well, and so do v_j and v_k, then it is likely that v_i and v_k may also achieve smooth communication and well collaborate as team members; and the transitivity applies as long as there exists a path between the pair of two individuals in G.

Based on the assumption above, we replace shortest path distance $sp(v_i, v_j)$ with the following definition, to define the communication cost between two individuals v_i and v_j as

$$d(v_i, v_j) = \frac{sp(v_i, v_j)}{length(v_i, v_j)}\ , \tag{3}$$

where $length(v_i, v_j)$ defines the number of edges involved in the path. For measuring the communication cost of a team X which forms subgraph G', we referred to the concept of average path length and made the modification on the four selected communication cost metrics as follows:

- $Cc - R'(X)$ is defined as the maximum $d(v_i, v_j)$ of any pair of members in X.
- $Cc - MST'(X)$ is defined as the cost of the minimum spanning tree on G' divided by $(|X| - 1)$ where $|X|$ stands for the size of team X.
- $Cc - SD'(X)$ is defined as the sum of all $d(v_i, v_j)$ for all pairs of members in X, divided by $(|X| * (|X| - 1)/2)$.
- $Cc - LD'(X)$ is defined as the sum of all $d(v_i, v_L)$ in X divided by $(|X| - 1)$, where v_L is the team leader and v_i is any other member.

Table 3. PCC between four modified communication cost metrics and team performance for observation data with 3623 records.

Communication cost metric	Pearson correlation coefficient
$Cc - R'$	-0.50132
$Cc - MST'$	-0.46810
$Cc - SD'$	-0.51658
$Cc - LD'$	-0.46006

Table 4. Statistics of regression analysis on modified communication cost metrics and team performance.

Communication cost metric	R Square	Significance F	Standard Error	Coefficient β_1	Intercept β_0
$Cc - R'$	0.251322	6.8E-230	1.388732	-0.51518	8.846803
$Cc - MST'$	0.219118	9.5E-197	1.418290	-0.69157	8.785556
$Cc - SD'$	0.266855	2.2E-246	1.374257	-0.74434	9.131815
$Cc - LD'$	0.211655	2.9E-189	1.425046	-0.68587	8.675663

4.2 Verifying the Improvement

Correlational analysis was performed under the same setup as chapter 3 to verify the feasibility of the modification. Results are shown in Table 3. While being negative for PCC of all four new metrics, the absolute value also increases. Following the conclusion drawn in chapter 3, we consider the new metrics more effective in measuring communication cost in the movie production case.

We have conducted a regression analysis between communication cost measured by new metrics and team performance to further support the verification. Communication cost measured by each of the new metrics is picked as the independent variable *x* (the predictor) while the corresponding team performance (assessed by movie ratings) is taken as the dependent variable *y* (the response) respectively. The results of correlational analysis using PCC suggest certain linear dependence of team performance on communication cost, therefore we built the following model from observation data, using the least square method:

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon \ , \tag{4}$$

where $\beta_0 + \beta_1 \cdot x$ is the deterministic part and ε is the residual (random error assumed normally distributed and independent of *x*). The statistics in Table 4 show the result of four groups of regression analysis.

Notice that for all four groups, the *p-value* (Significance F) < 0.0001, so we conclude that the regression model (Equation 4) is a good fit at significant level 0.0001. More importantly, the *R Square* shows the percentage of variance in team performance explained by the linear model. By comparing the *R Square* among these four metrics, we noticed the results in statistical analysis being consistent with those shown in Table 3, which indicate that $Cc - SD'$ may work better in measuring communication cost. The lower value of *Standard Error* (1.374257) also corresponds with the above result of comparison.

We may now conclude that our modified definitions of communication metrics are more effective comparing to the original design. Moreover, the new metric $Cc - SD'$ may be a better choice in terms of modeling team formation problem in the current case.

5 Conclusion

The presented study focuses on the team formation problem with concern of the collaboration factor. The general idea that effective collaboration would help achieve better team performance applies to group work in various domains. However, while related researches have contributed a variety of metrics and methods to model different communication structures within teams, there is a lack of investigation on the effectiveness of the proposed metrics. In this paper we conducted an empirical study to obtain better understanding to the problem, in which we referred to representative studies in the stream and selected four metrics for analysis. We used the real example of movie production teams, and utilized statistical analysis methods to reveal the potential correlation between communication cost and team performance. Results of analysis demonstrated that communication cost does have an adverse effect on team performance. Furthermore, we made improvements on the design of communication cost metrics based on the existed ones and verified the effectiveness in the same context of movie production teams.

Nevertheless, it is necessary to state that the results of analysis on those metrics may not apply to other cases and we do not intend to draw generalization from one single empirical study. There exist a lot of factors that may influence team performance other than collaboration, and they may vary in different domains. Still we expect the insights gained from the current empirical study would inspire researches on team formation problem to account for the effectiveness of modeling collaboration.

Further extension to the presented work may consider including more real-life examples for case study, such as DBLP, BibSonomy, etc. Comparative study among different cases might result in more valuable and solid conclusion. Also, in-depth statistical analysis methods could be applied to give more comprehensive understanding.

Acknowledgments. This work is supported by the National Natural Science Foundation of China under Grant No.61572539; the Research Foundation of Science and Technology Major Project in Guangdong Province under Grant Nos.2015B010106007, 2016B010110003; the Research Foundation of Science and Technology Plan Project in Guangdong Province under Grant No.2016B050502006; the Research Foundation of Science and Technology Plan Project in Guangzhou City under Grant Nos.2016201604030001, 201704020092. Information courtesy of IMDb (http://www.imdb.com). Used with permission.

References

1. Kumar, A., Dijkman, R. and Song, M., 2013. Optimal resource assignment in workflows for maximizing cooperation. In Business process management (pp. 235-250). Springer, Berlin, Heidelberg.
2. Liu, R. and Kumar, A., 2014, October. Impact of socio-technical network on process performance. In Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2014 International Conference on (pp. 243-252). IEEE.

3. Lappas, T., Liu, K. and Terzi, E., 2009, June. Finding a team of experts in social networks. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 467-476). ACM.
4. Turrentine, B., Calland, J.F., Adams, R., Shin, T. and Guerlain, S., 2003, October. Studying communication patterns during surgery. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 47, No. 12, pp. 1488-1492). Sage CA: Los Angeles, CA: SAGE Publications.
5. Kargar, M. and An, A., 2011, October. Discovering top-k teams of experts with/without a leader in social networks. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 985-994). ACM.
6. Juang, M.C., Huang, C.C. and Huang, J.L., 2013. Efficient algorithms for team formation with a leader in social networks. The Journal of Supercomputing, 66(2), pp.721-737.
7. Internet Movie Database, http://www.imdb.com/interfaces
8. Tjosvold, D., Moy, J. and Sasaki, S., 1999. Co-operative teamwork for service quality in East Asia. Managing Service Quality: An International Journal, 9(3), pp.209-216.
9. Kargar, M., An, A. and Zihayat, M., 2012. Efficient bi-objective team formation in social networks. Machine Learning and Knowledge Discovery in Databases, pp.483-498.
10. Majumder, A., Datta, S. and Naidu, K.V.M., 2012, August. Capacitated team formation problem on social networks. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1005-1013). ACM.
11. Lin, S., Luo, Z., Yu, Y. and Pan, M., 2013, September. Effective Team Formation in Workflow Process Context. In Cloud and Green Computing (CGC), 2013 Third International Conference on (pp. 508-513). IEEE.
12. Anagnostopoulos, A., Becchetti, L., Castillo, C., Gionis, A. and Leonardi, S., 2012, April. Online team formation in social networks. In Proceedings of the 21st international conference on World Wide Web (pp. 839-848). ACM.
13. Anagnostopoulos, A., Becchetti, L., Castillo, C., Gionis, A. and Leonardi, S., 2010, October. Power in unity: forming teams in large-scale community systems. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 599-608). ACM.
14. McDonald, D.W., 2003, April. Recommending collaboration with social networks: a comparative evaluation. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 593-600). ACM.
15. Wolf, T., Schröter, A., Damian, D., Panjer, L.D. and Nguyen, T.H., 2009. Mining task-based social networks to explore collaboration in software teams. IEEE Software, 26(1), pp.58-66.
16. Van Der Aalst, W.M., Reijers, H.A. and Song, M., 2005. Discovering social networks from event logs. Computer Supported Cooperative Work (CSCW), 14(6), pp.549-593.
17. Wang, X., Zhao, Z. and Ng, W., 2016. Ustf: A unified system of team formation. IEEE Transactions on Big Data, 2(1), pp.70-84.