Automatic Classification of Scientific Groups as Productive: An Approach based on Motif Analysis

Tanmoy Chakraborty*, Niloy Ganguly†, Animesh Mukherjee‡
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur, India – 721302
{*its_tanmoy,†niloy,‡animeshm}@cse.iitkgp.ernet.in

***Abstract*—One of the key aspects instrumental in the advancement of science relates to "team science," or in other words "group" collaborations. There have been extensive studies analyzing various statistical properties of collaborations of individual or pairs of authors. However, the number of studies pertaining to groups/teams of scientists working together is limited in number. In this paper, we set an objective to study the productivity of group collaborations where groups are represented as small substructures usually termed as *network motifs* in the literature. A preliminary observation is that star-like motifs have the largest productivity (defined as a function of citation count) followed by 4-cliques. We then introduce a bunch of features and study their individual relations with the productivity of a team. Building on these observations, we develop a supervised classification model that can automatically distinguish the highly productive teams from the low productive ones based on the set of identified features. The accuracy of the classification is 82% on an average for all the motifs with the accuracy reaching as high as 95% for 4-cliques. Finally, we present a detailed analysis of the time-transition behavior of different motifs along with some of the real world highly productive motifs found in our dataset. This empirical study is a first step toward the development of a full-fledged recommendation system that can predict how productive a team would be in the future.**

I. Introduction

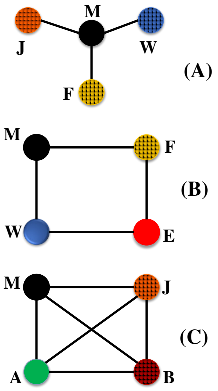
One of the key aspects of a scientific community is the prevalence of "team science" or group level collaborations. In fact, such group level collaborations constitute the building blocks of any collaboration network, i.e., a network where nodes represent authors and two authors are connected by an edge if they have co-authored one or more papers [1]. Now-a-days, collaboration among researchers seems to be increasing in popularity due to the increasing extent of "knowledge sharing" and cross-hybridization of multiple ideas [2]. One can leverage on the idea of local connectivity patterns of nodes, i.e., small groups/teams usually referred to as *network motifs* as a means for exploring the characteristic properties of a collaboration network since such recurrent local substructures often provide a crucial mesoscopic view at the intermediate scale between the whole network and the individual node.

The study of collaboration formation is an older research topic and started in parallel with the general research on collaboration networks [1] [3] [4] [5] [6]. All these analysis mostly concentrate on pair-wise collaboration between researchers. However, we identify that besides such one-to-one collaborations, the processes and outcomes of collaborative, team-based research, known as *group collaboration* where more than two researchers actively participate as a team in order to produce quality research can be extremely important and of significant interest to the scientific community. For example, Figure 1 shows three typical connectivity patterns centered around Mark Newman[1], a renowned British physicist at the University of Michigan. One can notice from the three patterns (Figures 1(a)-(c)) that even if Newman plays an important role in each case, the overall impact of each of these local groups (in terms of productivity as defined in section IV) varies significantly. This immediately indicates that there is a latent micro-dynamics governing the formation of different local substructures that needs to be investigated in order to understand the actual role of an individual within a team and to predict the fate of such groups in future.

Here, for the first time, we particularly investigate the group collaborations in a collaboration network in terms of "network motifs" [7] which are small subgraphs with a specific interaction pattern recurrently appearing in the network [8] [9] (see Figure 2). Note that, a group collaboration does not refer to the fact that all individuals are connected with each other; rather several combinations of a fixed set of individuals with different connectivities may form distinct group structures. Understanding such network motifs in a collaboration network has various utilities – (i) it can provide us an idea of the micro-level behavior of a group collaboration, (ii) the role of an individual in a group can be systematically investigated, (iii) success of different motifs over the years might enable a rising scientist to build new patterns of connectivity among her collaborators, (iv) network motifs which are mostly known as the functional blocks in biological science [8] might also help to unfold the functional role of network context around different individuals in the collaboration network.

The contributions of our work are manifold. We begin by defining a fundamental goodness measure of group collaborations – *productivity* (section IV). A simple analysis of productivity leads us to various interesting observations such as the star and the 4-clique motifs have relatively higher productivity than the rest of the lot; this observation is in sharp contrast with previous results reported by Krumov et al. [10]. Subsequently, we conduct a detailed investigation of a set of static features of motifs and try to draw correlation between productivity and these features (section V). The features that we consider are – (i) how long does a particular motif pattern take to form (construction time), (ii) the degree of heterogeneity of a group in terms of the research experience

[1] http://www-personal.umich.edu/~mejn/



Motifs	W_{av}	W_t
(A)	24.47	61.00
(B)	4.26	14.33
(C)	13.64	41.62

M:	Mark Newman
J:	Joshua Palay
W:	William Aiello
F:	Frank Thomson
E:	Eric J. Schwabe
A:	Abhiram G. Ramade
B:	Bruce M. Maggs

Fig. 1: (Color online) Three types of connectivity patters (motifs) centered around Mark Newman (M). Here, W_{av}, W_t correspond to the average productivity and the total productivity respectively as described in section IV. Note that, these examples are directly taken from our dataset.

of the constituent researchers (experience diversity), (iii) the variation of scientific impact of the constituent researchers (citation variance), (iv) the period of stability of a motif since its formation (recency) and (v) how long does a motif survive (longevity). As a second objective, we develop a classification model that can categorize high and low productive groups using the set of features mentioned above. The classification model shows reasonably high accuracy in classifying the motifs and the results are remarkably good for the 4-clique motifs (95% overall accuracy) (section VI). Then, we present a detailed discussion of our results by showing a dynamical characteristic of a motif – their time-transition behavior and the correlation of the same with the gain in productivity along with some real world examples of highly productive motifs (section VII). Finally, we conclude the paper by mentioning some interesting insights of this study and some immediate future directions (section VIII).

II. Dataset and Network Construction

We have used the dataset of the computer science domain used by Chakraborty et al. [11]. The dataset contains the name of the research paper, index of the paper, its author(s), the year of publication, the publication venue, the list of research papers the given paper cites and (in some cases) the abstract of the papers. In order to make the data suitable for our experiments, we extract only those entries which contain the information about the paper index, the title, author(s), the year of publication and the citations. Some of the general information pertaining to the filtered dataset of computer science are presented in Table I.

For the author name disambiguation, we use "RankMatch" algorithm[2] proposed by Liu et al. [12]. There are a couple of reasons behind adopting this algorithm. First of all, it is a completely unsupervised approach which is required in our study. In addition, the algorithm has been proved to be effective for the same types of scientific dataset [12]. The algorithm first assigns an unique index ID to all the author names present in the dataset. Then it follows a two-step strategy. (i) For each indexing author ID, it tries to pull out all the authors whose author names are possible variations of the indexing author name. To come up with the pool, it takes into account a number of cases where names can mutate or be disturbed. (ii) In the second step, it trims the candidate pool based on authors' publication features. Examples of publication features include co-authorship network, publication venues, years, title words. These features turn out to be discriminative for identifying real duplicates from the candidate pool. The number of authors after author name disambiguation is shown in Table I.

TABLE I: General information of the filtered dataset of the computer science domain.

Number of valid indices of papers	702,973
Number of authors before author name disambiguation	501,425
Number of authors after author name disambiguation	495,311
Average number of papers by an author	3.52
Average number of authors per paper	2.609
Time interval of the used dataset	1980 – 2005

The next task is to construct the collaboration network from the tagged dataset. Formally, a collaboration network is defined as a graph $G =< V, E >$ where each node $v_i \in V$ represents a researcher and an undirected edge e_{ij} between v_i and v_j is drawn if two researchers represented by v_i and v_j collaborate at least once via publishing a paper. From the above dataset, an overall collaboration network G has been constructed with researchers representing nodes and undirected edges representing collaborations between two researchers. As a new researcher starts her research career, she may enter or leave different collaborations. We track the changes in collaborations for a particular researcher over her entire research career. For this purpose, we analyze the collaboration network G_i composed of all nodes and edges between t_0 and t_i where t_0 is the earliest year present in the dataset. We call each such G_i a "snapshot" throughout the rest of the paper. Thus in each snapshot, all the edges of a collaboration since the beginning of the career of an author is present. In other words, we do not consider the deletion of a collaboration edge and if an edge is ever established it continues to be present in all the subsequent G_is constructed. Further note that, from our data it is possible to obtain a list of characterizing features of an author node as well as a collaboration edge – the total number of citations received by the authors, the year when an author makes her first/last publication, the number of co-citations obtained by an author pair and the year when an author pair make their first/last joint publication.

III. Motif Detection in Collaboration Network

In order to detect motifs, we use the "FANMOD"[3] proposed by Wernicke and Rasche [13] which is a tool for fast network motif detection. It relies on recently developed algorithms to improve the efficiency of this task by some orders of magnitude compared to existing tools [14]. FANMOD can detect network motifs up to a size of eight vertices using a novel algorithm called RAND-ESU [15][4]. We detect all

[2] The code is publicly available in https://github.com/remenberl/KDDCup2013.

[3] http://www.minet.uni-jena.de/~wernicke/motifs/

[4] Since RAND-ESU is a randomized algorithm, it produces different number of motifs in different iterations. Therefore after 50 iterations, we use the results obtained from that iteration which produces the maximum number of motifs.

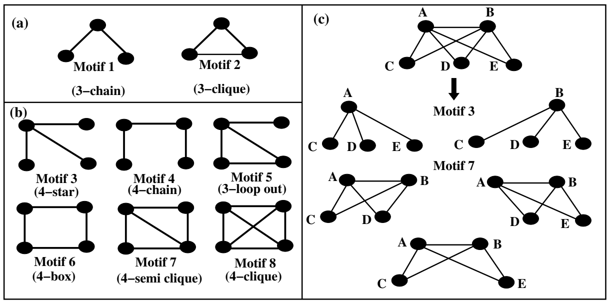


Fig. 2: The eight possible undirected (a) 3-node and (b) 4-node motifs with their standard names taken from the literature [7]. (c) Example of a local neighborhood structure in a collaboration network - 4-node motifs are extracted from the structure.

3-node and 4-node motifs from the overall collaboration graph (G) and each incrementally cumulating graph (G_i) constructed for every year. We restrict our analysis to 3- and 4-node motifs since the average number of authors per paper is found to be approximately 3 in our dataset (see Table I) We obtain two different combinations of 3-node motifs and six different combinations of 4-node motifs as shown in Figure 2 ((a) and (b)). For instance, in Figure 2(c), we obtain five different induced subgraphs (motifs) composed of four nodes for a example hypothetical network. Note that, FANMOD algorithm detects 3-node and 4-node motifs in two separate runs. Therefore, in the rest of the experiment, we analyze the 3-node and 4-node motifs separately. We have removed all such anomalous cases where the “longevity” of a motif – that is the difference in the number of years between the author pair who collaborated latest in the motif and the author pair who stopped collaborating earliest in the motif – is negative (discussed in further details in section V) which essentially indicates that such group collaboration never existed in the network. This filtration in turn deletes the invalid groups from the entire motif set. The motif distribution of the filtered collection of motifs in the overall collaboration graph G is shown in Figure 3(a). Chain motifs are found to be most prevalent. This result is also true for all year-wise subgraphs (G_i) as shown in Figure 3(b).

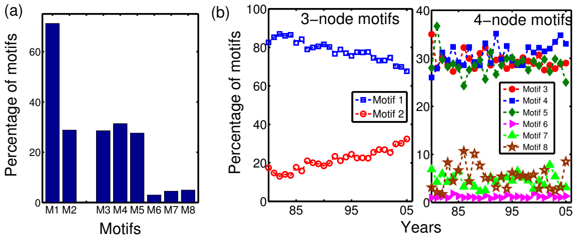


Fig. 3: (Color online) (a) Fraction of each type of 3-node motifs among all 3-node motifs and the fraction of each type of 4-node motifs among all 4-node motifs in the overall collaboration graph G (Mi stands for Motif i), and (b) their year-wise distributions.

IV. Measuring effectiveness of motifs

In this section, we measure the effectiveness of the motifs through the formulation of a fundamental quantifier of group collaborations: *productivity*. Since the number of papers published is not a quality metric, we define the productivity of a motif in terms of the average citation frequency per edge of all the involved publications. These citation frequencies serve as our surrogate measure for the impact of the publication. A crucial step is to convert the impact of publications into edge weights in the collaboration network. This conversion can be done in several different ways. We adopt two most effective measures proposed by Krumov et al. [10] for quantifying productivity of a motif.

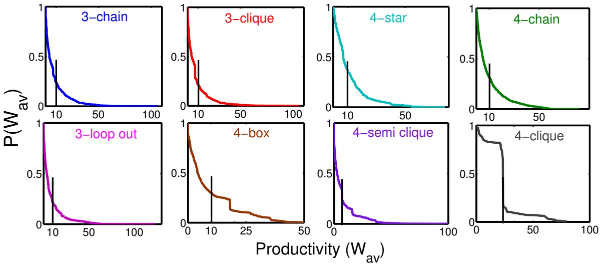


Fig. 4: (Color online) Cumulative distributions of W_{av} for all motifs. The vertical line in each frame indicates the cutoff based on which we develop a binary SVM classification model to distinguish productive motifs from the others.

For an edge e in the motif, let $P(e)$ denote the set of publications represented by e. For a publication p, $c(p)$ denotes the citation frequency of p. Then the *productivity* of a motif can be defined as follows:

$$W_t = \frac{1}{|E|} \sum_{e \in E} \sum_{p \in P(e)} c(p) \qquad (1)$$

where E is the set of edges in a motif. The subscript t is used to indicate the “total” productivity not normalized by the number of publications. Alternatively, if we wish to normalize with the number of publications then the equation can be rewritten as

$$W_{av} = \frac{1}{|E|} \sum_{e \in E} \frac{1}{|P(e)|} \sum_{p \in P(e)} c(p) \qquad (2)$$

It is not a priori clear which of the two measures defined above is the best way to define productivity since each has its own justification. Therefore, we use both the measures separately while calculating productivity of a motif in the rest of the experiments.

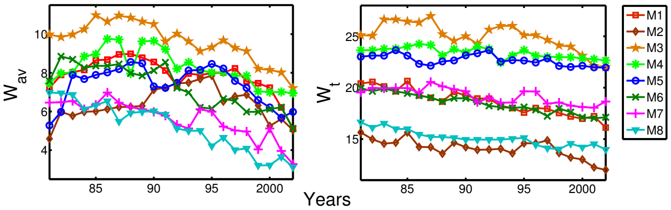


Fig. 5: (Color online) Year-wise productivity distribution per motifs. Mi stands for Motif i.

Distribution of productivity: We plot the distribution of productivity for all the motifs in Figure 4. We observe that both the productivity measures follow similar distribution.

Therefore, for the sake of clarity and conciseness, we only plot the distribution of W_{av} in Figure 4. In each plot, we draw a vertical line to indicate the threshold that marks a high-productive motif (mostly concentrated in the tail of the distribution) from the rest of the lot. We further plot the year-wise behavior of the productivity of different motifs in Figure 5. This indicates that the star motif (Motif 3) and the 4-clique motif (Motif 8) have a relatively higher productivity. The reason for the high productivity of these two motifs can be intuitively explained as follows – while for the star motif the central node is possibly representative of a very important scientist and a majority of the productivity of such a star motif can be attributed to this "center of power"; the 4-clique on the other hand is the ultimate "stable point of attraction" for all the other structures. Note that, the concentration of 4-cliques is not very high (see Figure 2) in the system which indicates that it takes long enough (due to "add-edge one" behavior as we shall see later) before other structures can finally land up to this highly productive penultimate configuration. Another important point that is reflected in the year-wise analysis is that while the values of W_{av} for all motifs start coinciding in the years after 2000, the same is not true for W_t. This is possibly because there is an exponential increase in the total number of publications and the normalization of the citation counts with such "astronomic" number of publications that forces the W_{av} of all the different motifs to coincide. These observations are in sharp contrast with previous results reported by Krumov et al. [10] where they report that the box motif has the maximum productivity as compared to others.

V. Motif characteristics

In this section, we first identify a set of discriminative features that could be attributed to the characteristics of a group (i.e., a network motif). All the distinctive features of group collaborations are derived from the characteristics of the constituent authors. We also analyze the correlation of the following features with productivity for all the motifs.

A. Construction Time (CT)

Since each individual edge in a motif indicates an one-to-one collaboration, it is associated with a year, the year when two collaborators published their first joint paper. Therefore, an edge is created by the first publication of the authors constituting this edge. For an occurrence of a motif, the construction time is the time between the earliest and the latest year of creation of the edges that constitute the motif. Formally, the *Construction Time (CT)* of a motif M is defined as: $CT(M) = Max(Cr(e_i)) - Min(Cr(e_i)) + 1, \forall e_i \in M$, where $Cr(e_i)$ = year of creation of edge e_i ($e_i \in M$). For example, if a 3-node motif M is constructed from three edges e_1, e_2 and e_3, and $Cr(e_1)$ = 1972, $Cr(e_2)$ = 1973, $Cr(e_3)$ = 1974; then the construction time of M is $CT(M)$ = (1974 - 1972) + 1 = 3 years.

We intend to examine whether the construction time has any effect on the productivity of a motif. The frames in the first column of Figure 6 show the average productivity of all occurrences of a particular motif that have the same construction time. The curves show that the construction time does not bear a very strong correlation with the productivity for any of the motifs. This indicates that the time required for a group to come to existence does not, in general, strongly determine the overall quality of the group.

B. Experience Diversity (ED)

The group collaborations can be categorized based on the duration of research experience of the constituent researchers forming the group. For instance, a group comprising a supervisor and her students is different from a group containing contemporary researchers. Note that, by the term "research experience" of a researcher, we mean the time difference from the earliest year when she published her first paper to the present time. The more the diversity (variance) of the research experience of the constituent collaborators in a motif, the more the motif indicates a group led by the senior researcher(s) with young fellows (e.g., supervisor-student group). We would like to check whether there is an effect of overall experience diversity of a group on productivity. The frames in the second column of Figure 6 show the productivity of all the motifs arranged in various ranges of experience diversity. We observe that the productivity (for both the measures) decreases with the increase of experience diversity of a group. From this, we might conclude that the groups comprising peer researchers of similar experience are much more productive compared to the groups led by a single experienced researcher.

C. Citation Variance (CV)

Another important feature that makes a researcher recognized in the scientific community is the average number of citations received by the papers she has published. A long span of research experience of an author may not indicate high number of average citations per paper she published. Here, for a researcher, we extract the overall number of citations (normalized by the number of papers) received by that researcher. Then similar to the earlier experiment, we find out the variance of the normalized citation counts of all constituent researches in a motif. Essentially, we are interested to see how the citation variance drives the productivity of a group collaboration, i.e., are the groups containing all highly cited researchers superior than the less cited groups? In the two frames of the third column in Figure 6, we observe that except in star motif (Motif 3), all the other motifs show a consistent pattern that the average productivity increases with the increase of citation variance. This result is markedly in contrast to the earlier results shown for experience diversity. Therefore, these two results imply that experience diversity and citation variance are not at all correlated when measuring with respect to the productivity of a motif. We shall discuss this in more details in section VI.

D. Recency (RC)

As the citation counts accumulate over time it is important to have a measure of the age of a group and observe its relationship with the productivity metric. The recency of a motif indirectly indicates the amount of time the motif is staying in the system without getting converted to a different motif. Note that, the clique motifs (M2 and M8) cannot get converted as we do not consider deletion. We study as a feature the number of years since the motif was fully created. In order to find out the recency of a motif, we map the motifs between two consecutive years and measure how long the motif under inspection is stable without any further edge addition (see section VII-A). We expect that the longer a group (motif) stays,

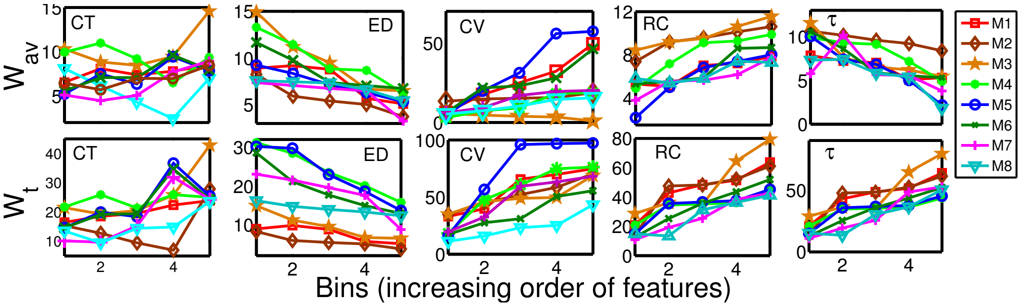


Fig. 6: (Color online) Variation of productivity (both W_{av} and W_t) of motifs with different characteristic features mentioned within each frame. All motif instances are grouped into five bins (1: low, 5: high) according to the values of the corresponding characteristic features (Mi stands for Motif i).

the more citations it might receive. The frames in the forth column of Figure 6 show that productivity of all the motifs increases with the increase in the stabilization time.

E. Longevity

Longevity of a group collaboration is also one of the important characteristics. We define the longevity of a motif as the number of years between the commencement of the last collaboration and the termination of one of the collaborations. For instance, let us assume a 3-chain motif M having edges e_1, e_2 and e_3. Each individual edge denotes an one-to-one collaboration. Let us denote the creation times (when two end-researchers of an edge published their first paper together) of these three collaborations by $Cr(e_1)$, $Cr(e_2)$ and $Cr(e_3)$ (say, $Cr(e_2) \geqslant Cr(e_1), Cr(e_3)$) respectively and the time when two end-researchers of an edge published their last paper together of these three collaborations by $Dl(e_1)$, $Dl(e_2)$ and $Dl(e_3)$ (say, $Dl(e_1) \leqslant Dl(e_2), Dl(e_3)$) respectively. Then the longevity of M is ($Dl(e_1)$ - $Cr(e_2)$) +1 (we also consider the year when the last edge has been created). Formally, the *longevity* (τ) of a motif M is defined by the following equation: $\tau(M) = min(Dl(e_i)) - max(Cr(e_i)) + 1$, $\forall e_i \in M$, where $Cr(e_i)$ and $Dl(e_i)$ denote the creation and the deletion years of the edge e_i respectively. For example, if a 3-node motif M is constructed by three edges e_1, e_2 and e_3, and $Cr(e_1)$ = 1972, $Cr(e_2)$ = 1973, $Cr(e_3)$ = 1974, $Dl(e_1)$ = 1976, $Dl(e_2)$ = 1979 and $Dl(e_3)$ = 1984; then according to the equation, the longevity of M is $\tau(M)$ = (1976 - 1974) + 1 = 3 years. Note that, it may happen that τ becomes negative for a certain motif when the motif contains such an edge which is created after the year when one of the edges of that motif has already been destroyed. As mentioned in section III, we completely ignore such motifs in all our experiments. Surprisingly, in Figure 6 (last column) we notice that although W_{av} decreases with the increase of longevity, W_t increases significantly with longevity for all the motifs. This is probably an indication of the saturation of productivity of an existing collaboration with the increase of longevity, though they accumulate significant amount of citations for their published papers. Note that, recency and longevity are two completely independent measures, where the former is calculated based on the motif transition behavior, the latter is measured based on the time stamp associated with each edge of a motif.

VI. Classification Model

In this section, we discuss a classification model that can help classify the motifs based on their productivity measure. It takes into account a set of discriminating features as discussed in section V: construction time (CT), experience diversity (ED), citation variance (CV), recency (RC) and longevity (τ).

A. Feature correlations

Before entering into the detailed description of the model, we perform a systematic analysis of the correlations between the features in order to identify if any of the features is fully determined by some other feature(s) and thus may be dispensed. For this, we calculate the Pearson correlation among the features and plot them in a heat map in Figure 7. We observe maximum correlation between recency and longevity (0.31), followed by recency and construction time (0.21). The highest negatively-correlated pair is recency and experience diversity (-0.25), followed by citation variance and recency (-0.12). Most of the correlations among the pairs of features are very small or negative which imply that the feature set is highly discriminative and negatively-correlated. Note that, as we do not observe any of the features to be highly related (correlation of the order of 0.9 or more) to any other it is not possible to dispense some of them in lieu of the other. Therefore, we use all the features in the subsequent analysis and the classification model made in the rest of this section.

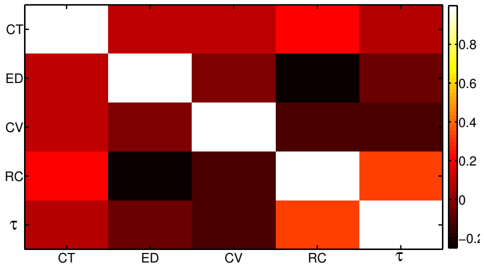


Fig. 7: (Color online) Heat map depicting the correlations among different features.

B. Evaluation metrics

To evaluate the performance of a binary-classification model, one can simply measure the overall accuracy of the system in comparison to the gold-standard dataset. The *Overall Accuracy* (OA) can be defined as follows:

$$OA = \frac{Number\ of\ correct\ classifications}{Total\ number\ of\ samples} \tag{3}$$

However, measuring only the OA may not properly indicate the true performance of the system, especially when the population on which the system is evaluated is biased towards a single class. Therefore, in order to measure the performance of the system at a more granular level, we also estimate the following metrics along with the OA:

$$Sensitivity(R^{+}) = \frac{Correctly\ classifed\ positive\ samples}{True\ positive\ samples} \tag{4}$$

$$Specificity(R^{-}) = \frac{Correctly\ classifed\ negative\ samples}{True\ neagtive\ samples} \tag{5}$$

$$PositivePrediction(P^{+}) = \frac{Correctly\ classified\ positive\ samples}{Positive\ classified\ samples} \tag{6}$$

$$NegativePrediction(P^{-}) = \frac{Correctly\ classified\ negative\ samples}{Negative\ classified\ samples} \tag{7}$$

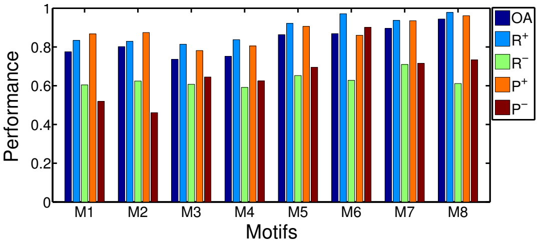


Fig. 8: (Color online) Performance of SVM model to classify motifs based on the productivity measure. Mi stands for Motif i.

C. Classification model based on productivity

This model is a binary classifier that tries to classify the motifs based on their productivity. To decide the cutoff among the spectrum of productivity values of motifs, we draw the distributions of W_{av} and W_t. We observe that both the distributions follow similar patters for all the motifs. Therefore, we only consider the distribution of W_{av} as shown in Figure 4 to decide the threshold. We observe that in most of the cases, the first dipping of the distribution of W_{av} occurs at the value of 10 in the x-axis of Figure 4. Therefore, the threshold is decided to be 10, i.e., the motifs having W_{av} <10 are considered as "low-productive" (positive class) and the rest as "high-productive" (negative class). From Figure 4, it is apparent that the population is highly biased towards the positive class. Here, we use Support Vector Machine (SVM) [16] as a supervised machine learning model to classify the motifs. For training and classification phases of SVM, we use YamCha[5] toolkit and TinySVM-0.075[6] classifier respectively with binary decision method and a linear kernel. We adopt a 10-fold cross validation technique where the whole population is divided into 10 chunks. We perform 50 different iterations and in each iteration, nine of them are randomly sampled out for training purpose and the rest one for testing. The performance of the classifier is measured for each of the motifs separately and pictorially depicted in Figure 8. On an average, this model shows nearly 82% accuracy for all the motifs. Here, the model more accurately classifies 4-clique motifs based on productivity (OA=0.95, R^{+}=0.98, R^{-}=0.62, P^{+}=0.96 and P^{-}=0.74) which is followed by semi-clique and box motifs. This result immediately shows that 4-cliques have a markedly different behavior as was also observed in the previous sections. Since they represent the penultimate configuration, the accuracy of the model should be the best for them and indeed so is the case. This again clearly justifies the significance of the use of motifs in this entire study as opposed to any other form of structural analysis. In addition, we observe that while R^{-} is greater than 60% throughout, P^{-} for the 4-node motifs is greater than 60% which is again a good achievement of the model.

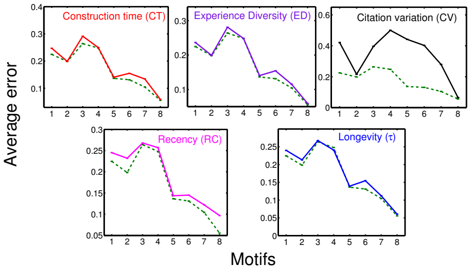


Fig. 9: (Color online) Error analysis for the classification model. Each frame shows the average error of the model when removing the corresponding feature mentioned in the frame. The broken green line depicts the average error when using all features. The number i in x-axis stands for Motif i.

Error analysis: We systematically analyze the significance of the features used in this model by dropping each of them in isolation and measure the performance of the model. Figure 9 displays the error that occurs due to omission of each feature. Here, while four of the features, namely construction time, experience diversity, recency and longevity seem to be only marginally important in the classification model, the citation variance (CV) turns out to be an extremely important classification feature for all the motifs. In particular, for 4-chain motifs (Motif 4) and 3-loop out motifs (Motif 5), dropping the citation variance can degrade the performance of the model nearly by three times. We also observe that it has highest correlation with the productivity measures (Pearson correlation of 0.29 and 0.39 with W_{av} and W_t respectively) compared to the other features. Therefore for a deeper analysis, we use only CV as a feature in the SVM model and measure the accuracy. Interestingly, we observe that though CV has highest correlation with the productivity measure and dropping it from the classification model causes maximum decrease in accuracy, keeping only CV in the model results in 62% average accuracy for all the motifs, which is quite low compared to the combined effect (82% average accuracy). Further analysis on the absolute values of the citation variance for all the motifs reveals that alone is not sufficient to determine proper discriminative boundaries corresponding to the two classes of productivity (high/low). Further, to examine the combined effect of the feature set, we include each feature one at a time along with CV in the SVM model in the following order (decreasing order of average error obtained from Figure 9): RC, CT,

[5] http://chasen.org/~taku/software/yamcha/

[6] http://chasen.org/~taku/software/TinySVM/

ED, τ. Then we measure the average accuracy for each addition and obtain accuracy values as 76%, 80%, 81%, 82% according to the ordered sequence mentioned above. Here, one can clearly notice that the combined effect also follows the same ordering as mentioned earlier along with the maximum gain obtained due to the addition of RC with CV. However, longevity seems to be less effective for classification model. This analysis indeed reflects the importance of individual features for distinguishing productive groups from the others.

VII. Discussion

In this section, we first discuss an important dynamical characteristics of motifs – their time-transition behavior and the effect of the time transition in the overall gain/loss of productivity. Then we further look back into the entire population of motifs and put forward some real examples of motifs comprising renowned computer science scientists found in our dataset to infer some interesting insights regarding group collaboration.

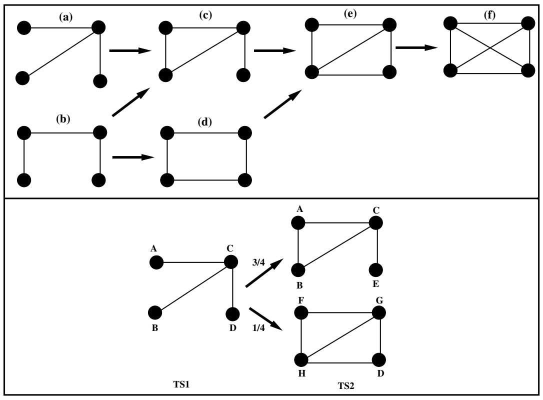


Fig. 10: Transition of 4-node motifs after adding a single edge (upper) and the toy example of mapping motif(s) across successive timestamps (TS) (lower).

A. Motif transition

As mentioned earlier, one of the primary objectives of our study is to analyze the motif transition over the time-periods that indicates the propensity of each motif to metamorphose to another. We have already mentioned the use of motif transition earlier when describing the recency of a motif. In a time-varying environment, if a single edge is added to a motif in each pass keeping the number of nodes constant, the structure of the motif changes into another form. For instance, addition of an edge can convert a 3-chain to a 3-clique. For 4-node motifs, the process follows a little complicated dynamics as shown in Figure 10 (upper). For instance, addition of a *single edge* in the system one at a time can lead to any of the following three paths (or the sub-paths): $(a) \rightarrow (c) \rightarrow (e) \rightarrow (f)$, $(b) \rightarrow (d) \rightarrow (e) \rightarrow (f)$ and $(b) \rightarrow (c) \rightarrow (e) \rightarrow (f)$. However, in real-world scenario, it can be possible that more than one edge get added between two consecutive timestamps.

We extract motifs from each of the year-wise graphs G_i. Now the next task is to map each motif in year t_i to one of the motifs in year t_{i+1}. Instead of one-to-one mapping, we adopt a one-to-many functional mapping technique shown in Figure 10 (lower). Here, if n nodes in a motif M at time t_i are divided between two motifs (say, M_1 and M_2) at t_{i+1} keeping m and $(n-m)$ nodes of M respectively, then we consider $\frac{m}{n}$ fraction of M is transformed into M_1 and rest $\frac{n-m}{n}$ of M is transformed into M_2. In this way, we compute the fraction of changes of one motif to others across all time transitions present in our dataset. Figure 11(a) shows this fraction (in %) for all the motifs. For instance, Motif 3 is transformed into Motif 5, Motif 7 and Motif 8 in 72.26%, 12.56% and 15.18% of overall transformations respectively. One important observation is that, most of the motif transitions show a similar behavior that they usually follow "add-edge one" behavior discussed in Figure 10 (upper), i.e., the fraction of transitions of one motif to the other motif(s) due to the addition of a single edge is higher than the fraction of transitions to other motif(s) through the addition of multiple edges. These results imply that the dynamics of group formation is usually not an arbitrary process, rather it evolves in a steady and systematic fashion with single edge addition in each transition.

We further study the cost of motif transitions in terms of the gain/loss of productivity. We define the *gain* of productivity (ΔW) due to motif transition as follows: $\Delta W = \frac{W_{new} - W_{old}}{W_{old}}$ (W can be replaced by W_{av} or W_t). Figure 11(b) shows that in all the transitions, the gain in productivity is positive when the final structure is the 4-clique (Motif 8). This again corroborates that 4-clique acts as the *final reservoir* for all the other structures, and therefore, the evolution is driven towards this structure. Another interesting observation here is that the productivity increases when a star motif gets converted to a clique motif, although in general star motif is more productive than clique motif (see Figure 5). However the chance of this conversion is rare (see Figure 11); hence in most cases clique motif appears after passing through several other intermediate motif configurations subsequently decreasing the productivity.

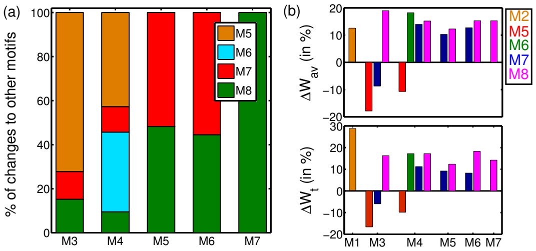


Fig. 11: (Color online) (a) Fraction (in %) of changes of one motif into the others across all time transitions and (b) gain in productivity due to motif transition (Mi stands for Motif i).

B. Analysis of productive motifs

Here we refer to some real world highly productive motifs found in our dataset and provide some interesting results for those motifs.

- **Jon Kleinberg's collaborations:** We find a star motif comprising Y. Rabani, E. Tardos, J. Kleinberg, F.T. Leighton (where J. Kleinberg[7] is placed at the center of the motif) to be the most productive group in our dataset in terms of W_{av}. However, the construction time of this group is 2 years and the experience diversity is also high. This group lasted around 10 years which is reasonably high in our dataset. On the other hand, if we observe another star motif constituting

[7] http://www.cs.cornell.edu/home/kleinber/

A. Aggarwal, M. Charikar, D. Williamson in three peripheral nodes and centered around J. Kleinberg, we get significantly different statistics. Although the construction time is similar to the earlier one, its productivity and longevity are very less. As expected, we observe that the second motif gets converted to a 4-clique (comprising A. Aggarwal, M. Charikar, J. Kleinberg and others) and a 3-loop out (comprising D. Williamson and others) in order to gain productivity.

- **Jiawei Han's collaborations:** We have noticed in section V that a 4-chain tends to get converted to a 4-clique motif more often in order to gain high productivity. A prominent evidence found in our dataset is the 4-chain comprising Jiawei Han[8], Yongjian Fu, Zhaohui Xie, Wei Wang (where three collaboration edges are formed between first-second, second-third and third-forth authors sequentially in order). This motif lasted for 3 years before augmenting three other edges to form a 4-clique, and this transformation produces a gain in 12% of W_{av} and 15% of W_t.
- **Michael I. Jordan's collaborations:** Interestingly, we observe that Michael I. Jordan[9] is present in star motifs maximum number of times and the recency of those motifs is also very small in comparison to the other star motifs. However, the 3 clique motif containing David M. Blei, Andrew Y. Ng, Michael I. Jordan seems to be the most productive in our dataset in terms of W_t. A deeper look into this collaboration reveals that this motif gets maximum citations due to the famous paper on "Latent Dirichlet Allocation" [17].
- **James Allan's collaborations:** Similar to the earlier observations, James Allan[10] is found to occur in maximum number of times in 4-cliques. However, maximum productivity is observed for the star motifs comprising J. Allan., J. Callan, W. B. Croft, and M. Hirsch centered around J. Allan. This motif lasted 8 years before converting to a 4 clique. However, this transformation achieves very less overall gain in W_{av} (2%) and W_t (4.5%). The maximum productive 4-clique motif among his collaborations constitutes J. Carbonell, G. Doddington and J. Yamron along with him.
- **Nicholas R. Jennings's collaborations:** A typical pattern found in most of the motifs centered around Nicholas R. Jennings[11] is that their experience diversity is quite high even if these motifs gain significant higher productivity. This is counterintuitive to our earlier observation in section V that the groups with high experience diversity tend to be less productive. The motif set constituting N. R. Jennings is mostly dominated by 4 cliques followed by the star motifs. However, the most long-lived motif centered around him constitutes K. P. Sycara, M. P. Georgeff and M. Wooldridge in the periphery that lasted for 5 years.

VIII. Conclusions

In this work, we showed that in the collaboration network, motifs have significant potential to unfold the underlying dynamical behavior of scientific teams. We proposed a set of characteristic features to describe such motifs and related them to productivity, a fundamental goodness measure of a group collaboration. We conclude the paper mentioning few interesting outcomes and some immediate future directions as follows: (i) we observe that the star and the 4-clique motifs are highly productive, (ii) the characteristics of a group collaboration can suitably be explained in terms of a set of distinctive features of the constituent researchers, (iii) in real-world, the transition of motifs over the successive time steps is usually not abrupt, rather it systematically follows "add-edge one" mechanism, (iv) transition to a 4-clique produces the largest gain in productivity for all the motifs, (v) the characteristic features of a group collaboration quite efficiently classify network motifs based on the productivity with the best classification obtained for the 4-clique motifs.

We believe that a stronger connection between the motif patterns and the underlying elementary processes in the system (selecting authors for a publication, selecting articles to be cited within a publication) can be achieved via generative minimal models [10]. The current analysis might also allow us to forecast the number of citations that an author/collaboration could possibly acquire in future thus leading to the design principles of an efficient recommendation system.

References

[1] M. Newman, "Coauthorship networks and patterns of scientific collaboration," *PNAS*, vol. 101, pp. 5200–5205, 2004.

[2] J. Huang, Z. Zhuang, J. Li, and C. L. Giles, "Collaboration over time: characterizing and modeling network evolution," in *WSDM*. New York, NY, USA: ACM, 2008, pp. 107–116.

[3] Y. Ding, "Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks," *J. Informetrics*, vol. 5, no. 1, pp. 187–203, 2011.

[4] L. Kronegger, F. Mali, A. Ferligoj, and P. Doreian, "Collaboration structures in slovenian scientific communities," *Scientometrics*, vol. 90, no. 2, pp. 631–647, Feb. 2012.

[5] A. Gazni, C. R. Sugimoto, and F. Didegah, "Mapping world scientific collaboration: Authors, institutions, and countries," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 2, pp. 323–335, Feb. 2012.

[6] R. Kundra and H. Kretschmer, "A new model of scientific collaboration part 2. collaboration patterns in indian medicine," *Scientometrics*, vol. 46, no. 3, pp. 519–528, 1999.

[7] U. Alon, "Network motifs: theory and experimental approaches," *Nat. Rev. Genet.*, vol. 8, no. 6, pp. 450–461, 2007.

[8] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[9] S. Choobdar, P. Ribeiro, S. Bugla, and F. Silva, "Comparison of co-authorship networks across scientific fields using motifs," in *ASONAM*, pp. 147–152, 2012.

[10] L. Krumov, C. Fretter, M. Müller-Hannemann, K. Weihe, and M. Hütt, "Motifs in co-authorship networks and their relation to the impact of scientific publications," *EPJB*, vol. 84, no. 4, pp. 535–540, 2011.

[11] T. Chakraborty, S. Sikdar, V. Tammana, N. Ganguly, and A. Mukherjee, "Computer science fields as ground-truth communities: Their impact, rise and fall," in *ASONAM*, 2013, pp. 426 – 433.

[12] J. Liu, C. Wang, J. Y. Liu, K. H. Lei, and J. Han, "Ranking-based name matching for author disambiguation in bibliographic data," in *KDD Cup 2013 Workshop*, 2013.

[13] S. Wernicke and F. Rasche, "Fanmod: a tool for fast network motif detection," *Bioinformatics*, vol. 22, no. 9, pp. 1152–1153, May 2006.

[14] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics*, vol. 20, no. 11, pp. 1746–1758, Jul. 2004.

[15] S. Wernicke, "A faster algorithm for detecting network motifs," in *WABI*. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 165–177.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[8] http://www.cs.uiuc.edu/~hanj/

[9] http://www.cs.berkeley.edu/~jordan/

[10] http://ciir.cs.umass.edu/~allan/

[11] http://users.ecs.soton.ac.uk/nrj/