Soc. Netw. Anal. Min. (2015)5:16
DOI 10.1007/s13278-015-0255-3

ORIGINAL ARTICLE

An author is known by the context she keeps: significance of network motifs in scientific collaborations

Tanmoy Chakraborty[1] · Niloy Ganguly[1] · Animesh Mukherjee[1]

Received: 31 December 2014 / Revised: 5 May 2015 / Accepted: 13 May 2015

Abstract Collaboration networks are elegant representations for studying the dynamical processes that shape the scientific community. In this paper, we are particularly interested in studying the *local context* of a node in collaboration network that can help explain the behavior of an author as an individual within the group and a member along with the group. The best representation of such local contextual substructures in a collaboration network are "network motifs". In particular, we propose two fundamental goodness measures of such a group represented by a motif—*productivity* and *longevity*. We observe that while 4-semi clique motif, quite strikingly, shows highest longevity, the productivity of the 4-star and the 4-clique motifs is the largest among all the motifs. Based on the productivity distribution of the motifs, we propose a predictive model that successfully classifies the highly cited authors from the rest. Further, we study the characteristic features of motifs and show how they are related with the two goodness measures. Building on these observations, finally we propose two supervised classification models to predict, early in a researcher's career, how long the group where she belongs to will persist (longevity) and how much the group would be productive. Thus this empirical study sets the foundation principles of a recommendation system that would forecast how long lasting and productive a given collaboration could be in future.

T. Chakraborty was financially supported by Google India Ph.D. Fellowship.

✉ Tanmoy Chakraborty
its_tanmoy@cse.iitkgp.ernet.in

Niloy Ganguly
niloy@cse.iitkgp.ernet.in

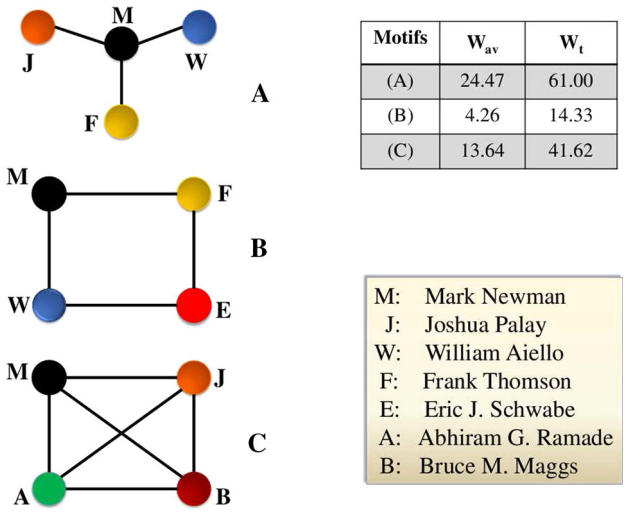
Animesh Mukherjee
animeshm@cse.iitkgp.ernet.in

[1] Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur 721302, India

1 Introduction

Coauthorship of a paper can be thought of as the documentation of a collaboration between two or more authors, and these collaborations form a "collaboration network" (a.k.a "coauthorship network") in which the network nodes represent authors, and two authors are connected by an edge if they have coauthored one or more papers (Newman 2004). Nowadays, collaboration among researchers has been increasingly popular through "knowledge sharing" and cross-hybridization of multiple ideas (Huang et al. 2008). The fine-grained assessment of collaboration network can lead to identifying local connectivity pattern of individuals to describe the network context to which they belong to (i.e., the local neighborhood of which the node under observation is a part of). One can use the idea of such local context of nodes as a means of exploring the entire collaboration network, since these local substructures not only describe the dynamics of collaboration patterns of an author over her entire research career but also provide a mesoscopic view at the intermediate scale between the whole network and the individual node. The significance of contextual information around a node has been successfully proved in understanding human languages (Dascal 1989), biological functional units (Prill et al. 2005) and topological structure of large complex systems (Hyun Yook et al. 2004). Here, for the first time, we particularly investigate the network context around an individual in a collaboration network as a combination of

Published online: 26 May 2015



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

Fig. 1 (Color online) Three types of connectivity patters (*motifs*) centered around Mark Newman (*M*). Here, W_{av} and W_t correspond to the average productivity and the total productivity, respectively, as described in Sect. 5. Note that these examples are directly taken from our dataset

different "network motifs" (Alon 2007) which are considered to be small subgraphs with a specific interaction pattern recurrently appearing in the network. For example, Fig. 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 (Fig. 1a–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 Sect. 5) varies significantly. This immediately indicates that there is a latent microdynamics governing the formation of different local substructures that need to be investigated to understand the actual role of an individual within a team and predict the fate of such groups in the future. In this research, we are primarily interested in such contextual information (in terms of network motifs Milo et al. 2002) around each individual node over different time periods that indeed explains the dynamics behind the changes in the collaboration profile.

We show how a systematic and rigorous analysis of such contextual information in terms of motifs can lead us to understand the micro-level behavior of authors (i) as an individual *within* the group and (ii) as a member of the collaboration, i.e., *along with* the group. Note that in the second case, all individuals may not be connected with each other, rather several combinations of a fixed set of individuals with different connectivities may form distinct group structures. While we address the first case briefly, the second case forms the major agenda of our current work.

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

We observe that the motif-based study presented in this paper is unique and remarkably unfolds certain individual and group-level characteristics of the authors that are not usually visible through direct statistical analysis. Note that the present research is an extension of our earlier observation where we showed that different patterns of scientific collaborations and can nicely be captured by network motifs (Chakraborty et al. 2014). Here, our primary observation is that the local context (represented by network motifs) of a node in a collaboration network can help explain the behavior of an author as an individual within the group and a member along with the group.

The contributions of our work are manifold. We begin by defining two fundamental goodness measures of a group—*productivity* and *longevity*. A simple analysis of these two metrics leads us to various interesting observations: (a) there are certain motif structures (e.g., 4-star, 4-clique) that have relatively higher productivity than the rest of the lot and (b) the more dense a motif is the longer does it last; the rate of new collaboration edges getting included into such dense motifs gets slower over time. We then investigate the behavior of an author within a group. In particular, we show that simple contextual information about the author allows us to nicely classify them in terms of the number of citations they receive and this is true even when there is an imbalance in the number of instances present in the two classes. It is important to mention here that such a motif-based study allows us to do the classification of even the rare class with an accuracy of 87 %. In a step further, these network contexts (motifs) in the collaboration network are assumed to be different representations of group collaborations. From this perspective, we investigate certain characteristic features of motifs to observe precise correlation between productivity/longevity and these features. The features that we considered 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 of the constituent researchers (experience diversity), (iii) the variation of scientific impact of the constituent researchers (citation variance) and (iv) the period of stability of a motif since its formation (recency). We observe that while productivity is not affected by (i), it is directly proportional to (iii) and (iv) and is inversely proportional to (ii). On the other hand, longevity is directly proportional to (i), inversely proportional to (ii) and (iii) and is not affected by (iv). In addition, we also investigate another important dynamical characteristic of a motif—their time-transition behavior and the correlation of the same with the gain in productivity/longevity. Here we observe that, in general, any transition in the structure of the motif causes an increase in the longevity, while a transition from any configuration to a 4-clique causes an increase in the productivity. As a final

objective, we develop two predictive models to suitably identify and predict the longevity and the productivity of a group (motif) based on the features discussed above. Both the models show reasonably high accuracy in predicting the longevity/productivity of a collaboration and the results are remarkably good for the 4-clique (overall accuracy of longevity prediction is 87 % and that of productivity prediction is 95 % in this case). Finally, we conduct a shallow level analysis of motif distribution in different fields of research and observe two distinct patters emerging from the distribution.

2 Related work

Research on collaboration network was started with the pioneer work of Newman (2001). After that, a large number of research works have been conducted on the statistical analysis of collaboration network (Ding 2011; Kronegger et al. 2012; Martinez-Romo et al. 2008; Said et al. 2008) and modeling collaboration network through simulations (Huang et al. 2008; Liu et al. 2012; Tambayong 2007). Similarly, few related researches on collaboration network, namely developing author-ranking scheme through "supportiveness" analysis (Han et al. 2009), ego-centric network analysis of collaboration network (Abbasi et al. 2012), classifying personal names through collaboration network (Biryukov 2008), discovering the relationship between authors and research domains (Hassan and Ichise 2009), understanding and modeling diverse scientific careers of researchers (Chakraborty et al. 2015) etc. have been conducted. Recently, Pan and Saramäki (2011) study the correlations between tie strengths and topology in networks of scientific collaboration and show that collaboration networks are very different from ordinary social networks.

In parallel, a group of researchers are engaged in experimenting another direction of research on collaboration network called "collaboration prediction". It includes nonparametric random effects model (Yu et al. 2009), maximum margin matrix factorization model (Rennie and Srebro 2005), proximity-based approach (Liben-Nowell and Kleinberg 2007), Supervised random work model (Backstrom and Leskovec 2011), etc. Recently, Krumov et al. (2011) conduct an experiment to demonstrate that motifs in the collaboration network represent different collaboration patterns and the success of individual authors or publications depends unexpectedly strongly on these intermediate scaled structures of collaboration networks. Choobdar et al. (2012) propose a motif-based approach to compare coauthorship networks across scientific fields. Similarly, Wu et al. (2012) classify Wikipedia articles using network motif counts and ratios. Shi et al. (2008) propose a scientific collaboration network evolution model based on motif emergence. Baras and Hovareshti (2011) develop a systems engineering-oriented approach to the design of networks of mobile autonomous systems. Yeang et al. (2012) modify and improve the method proposed by Milo et al. (2002) to detect significantly enriched motifs in both directed and undirected networks. They apply this method on the datasets of 18 networks including coauthorship network and show that the presence and absence of enriched motifs provide rich information regarding each type of network relations. Wu et al. (2008) use motifs to analyze the entire citation pattern of the journals indexed by CSTPC from year 2003 to 2006 and develop trend on journal networks. Lü and Zhou (2010) show that weak ties play a more significant role than the strong ties in the collaboration network using motif analysis.

To the best of our knowledge, this experiment is the first attempt whether motifs in collaboration networks are extensively studied over the years and their utilities are fully explored in different classification models. Moreover, tracking the motif transition over successive time periods not only explores different modes of individuals' collaboration patterns, but also aids in the overall productivity due to motif transition.

The rest of the paper is organized as follows. In Sect. 3, we give a brief description of our dataset. Following this, the technique to detect motifs is briefly narrated in Sect. 4. Then we present a detailed description of two goodness measures in Sect. 5. In Sect. 6, we study the author-level analysis of motifs and the author classification model. A detailed description of the set of features and their individual correlations with two goodness measures are shown in Sect. 7. Next in Sect. 8, we present two classification models to study the motifs at group level. Following this, an empirical correlation is drawn between two goodness measures in Sect. 9. Then a field-level analysis of motifs is shown in Sect. 10. Few real examples of productive motifs curated from our dataset are presented in Sect. 11. Finally, we conclude the paper with discussion and future work in Sect. 13.

3 Dataset and network construction

We have used the dataset of the computer science domain developed by Chakraborty et al. (2013). The dataset contains the name of the research paper, index of the paper, its author(s), the year of publication, the publication venue, the citations of a given paper and (in some cases) the abstract of the papers. The dataset is further distributed over 24 fields of the computer science domain (see Table 1).

To make the data suitable for our experiments, we extract only those entries which contain the information about

Table 1 Percentage of papers in various fields (with abbreviations) of the computer science domain

Fields	% of papers	Fields	% of papers
Artificial intelligence (AI)	12.64	Algorithms and theory (ALGO)	9.89
Networking (NETW)	9.41	Databases (DB)	5.18
Distributed systems (DIST)	4.66	Hardware and architecture (ARCH)	6.31
Software engineering (SE)	6.26	Machine learning (ML)	5.00
Scientific computing (SC)	5.73	Bioinformatics (BIO)	2.02
Human computer interaction (HCI)	2.88	Multimedia (MUL)	3.27
Graphics (GRP)	2.20	Computer vision (CV)	2.59
Data mining (DM)	2.47	Programming language (PL)	2.64
Security (SEC)	2.25	Information retrieval (IR)	1.96
Natural language and speech (NLP)	5.91	World wide web (WWW)	1.34
Education (EDU)	1.45	Operating systems (OS)	0.90
Embedded systems (EMB)	1.98	Simulation (SIM)	1.04

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 2.

For the author name disambiguation, we use "RankMatch" algorithm proposed by Liu et al. (2013).[2] 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 (Liu et al. 2013). The algorithm first assigns a 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, and 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 2.

The next task is to construct the collaboration network from the tagged dataset. Formally, a collaboration network is defined as a graph $G = \langle V, E \rangle$, 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 the 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 such 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 makes their first/last joint publication.

Table 2 General information of the filtered dataset of 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

4 Motif detection in collaboration network

To detect the motif, we use the "FANMOD" proposed by Wernicke and Rasche (2006) which is a tool for fast network motif detection.[3] It relies on recently developed algorithms to improve the efficiency of network motif

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

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

detection by some orders of magnitude over existing tools (Kashtan et al. 2004). FANMOD can detect network motifs up to a size of eight vertices using a novel algorithm called RAND-ESU (Wernicke 2005). We detect all 3-node and 4-node motifs from the overall collaboration graph (G) and each of the incrementally cumulating graphs (G_i) constructed for every year. We obtain two different combinations of 3-node motifs and six different combinations of 4-node motifs as shown in Fig. 2a and b. Note that, FANMOD algorithm detects 3-node and 4-node motifs in two separate runs. For instance, if we consider the network context depicted in Fig. 2c, we obtain various induced subgraphs composed of three and four nodes representing different motifs as follows: three 3-cliques ({A, B, C}, {A, B, D}, {A, B, E}), two star motifs ({C, A, D, E}, {C, B, D, E}), four 3-loop out motifs ({C, A, B, D}, {E, A, B, D}, {C, B, A, D}, {E, B, A, D}) and three 4-semi cliques ({A, B, E, D}, {C, A, B, D}, {D, A, B, E}). For this, in the rest of the experiment, we analyze the 3-node and 4-node motifs separately. In the rest of our analysis, 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 detail in Sect. 5). The motif distribution of the filtered collection of motifs in the overall collaboration graph G is shown in Fig. 3a. Chain motifs are found to be most prevalent. This result is also true for all year-wise subgraphs (G_i) as shown in Fig. 3b. Interestingly, we notice in Fig. 3b that the difference between the number of Motif 1 (3-chain) and Motif 2 (3-clique) gradually decreases over the years, i.e, most of the 3-chain motifs tend to shift to 3-clique motifs. This is perhaps due to the transitive relationship among the coauthors, which seems to become prominent over the years, as a result of which new edge gets attached with two

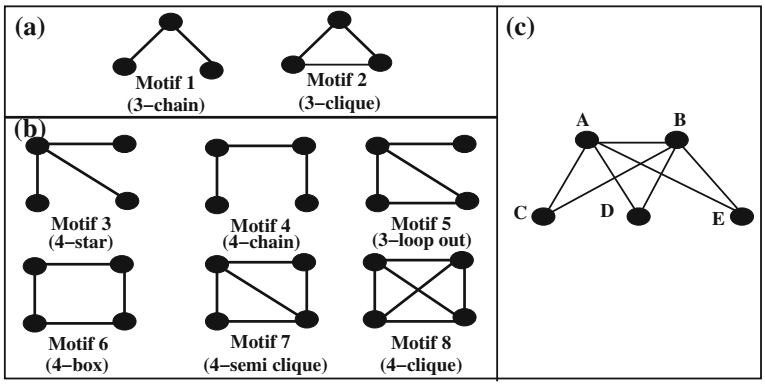


Fig. 2 The eight possible undirected **a** 3-node and **b** 4-node motifs with their standard names taken from the literature (Alon 2007). **c** Example of a local neighborhood structure in a collaboration network—motifs are extracted from the structure

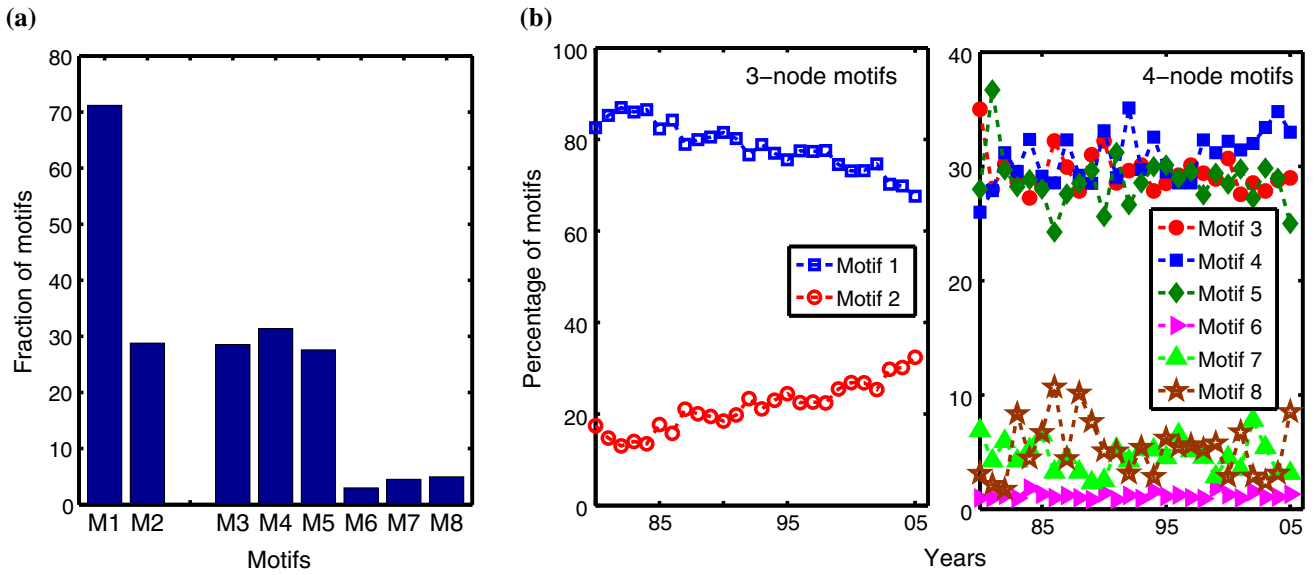


Fig. 3 (Color online) **a** Fraction of each type of 3-node motifs among all 3-node motifs and the fraction of each type of 4-node motifs among all 4-node motifs in the overall collaboration graph G (*Mi* stands for motif *i*) and **b** their year-wise distributions

coauthors who have at least one common coauthor. However, we observe that for 4-node motifs, the pattern is almost similar over the years.

5 Measuring effectiveness of motifs

In this section, we measure the effectiveness of the motifs through the formulation of two fundamental dimensions of collaborations: *productivity* and *longevity*. In particular, we quantify these two measures and report their distributions for different motifs.

5.1 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. (2011) for quantifying productivity of a motif.

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:

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

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

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

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

5.2 Distribution of productivity

We plot the distribution of productivity for all the motifs in Fig. 4. We observe that both the productivity measures follow a similar distribution. Therefore, for the sake of clarity and conciseness, we only plot the distribution of W_{av} in Fig. 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. The threshold is selected based on value in the x-axis corresponding to which the first dip is observed in the line for most of the motifs (here, the threshold is 10).

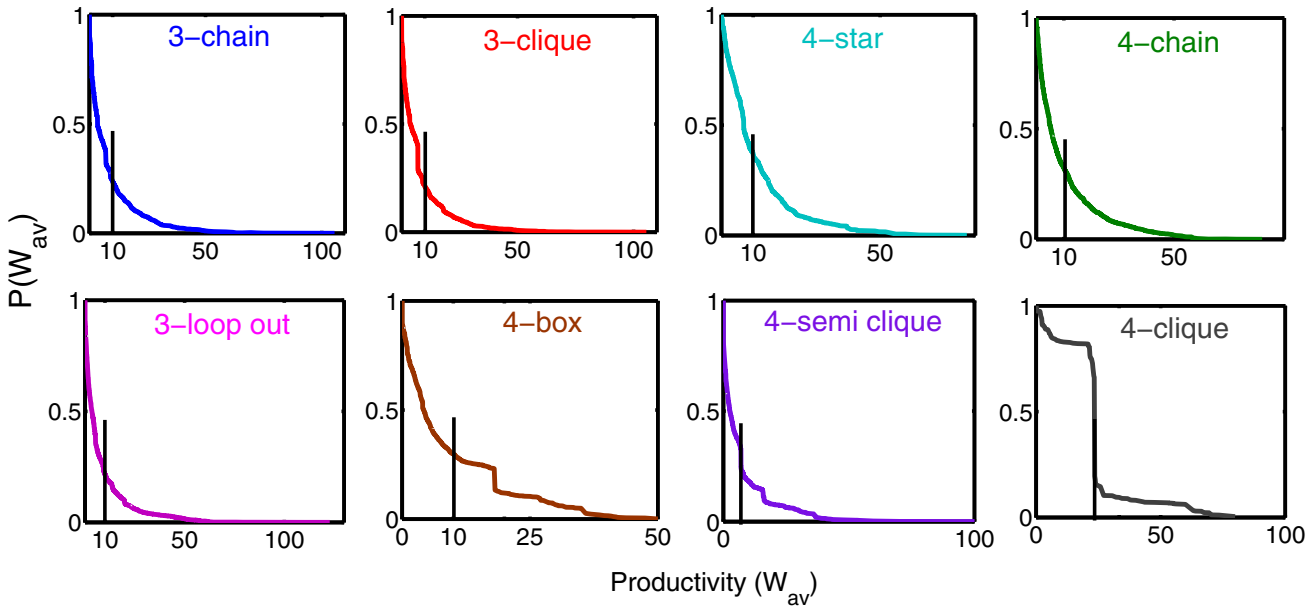


Fig. 4 (Color online) Distributions of W_{av} for all motifs. The *vertical line* in each frame indicates the cutoff based on which we develop a binary SVM classification model to predict the productivity of motifs

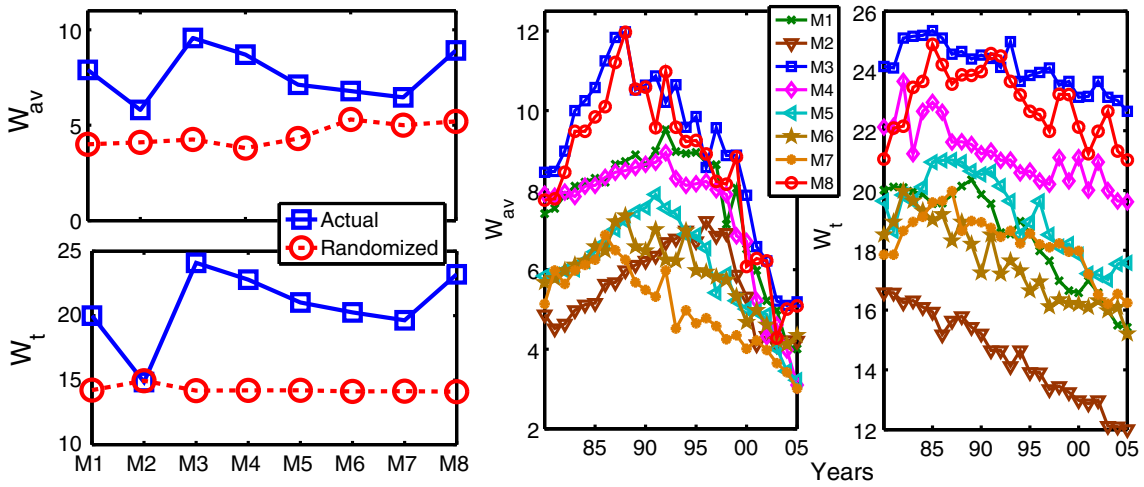


Fig. 5 (Color online) **a** The average productivity of the motifs compared to the null motifs according to the two measures of productivity and **b** year-wise productivity distribution per motifs (*Mi* stands for Motif *i*)

Moreover, we observe in Fig. 4 that the productivity distribution for Motif 8 (4-clique) is different from the rest in that there are a large number of 4-clique motifs which have productivity higher than the threshold, and there is a sharp decline at the threshold. The reason could be that in general the 4-clique motifs are highly productive (see Fig. 5). Therefore, most of the 4-cliques exhibit nearly similar higher productivity, which causes the line to be almost consistent in the initial part of the plot. However for the other motifs, since the number of highly productive motifs is less, the plot tends to decrease gradually. Figure 5a shows the average productivity (both W_t and W_{av}) of each motif. To show that the productivity of the motifs do not manifest arbitrarily, we randomize the citation frequencies of the publications and recompute the productivity of motifs in the *null-model* scenario of shuffled citation frequencies. A uniform distribution of these null-model edge weights across the motifs indicates a successful elimination of the residual influences (see Fig. 5a). It should be noted that in contrast to many case network analysis, we do not randomize the network architecture, but rather shuffle the weights of the edges. In this way, we do not consider the possible deviations of motif counts resulting from randomness, but only the effect the motifs have in shaping the dynamical output (here productivity) of the network. We observe in Fig. 5 that both the proposed measures are significantly different from the outcomes of the null model. To further test the robustness of our results, we plot the year-wise behavior of the productivity of different motifs in Fig. 5b. Both Fig. 5a and b indicate that the star motif (Motif 3) and the 4-clique motif (Motif 8) have a relatively higher productivity. Similar results are obtained when we measure the year-wise productivity—the star and the 4-clique motifs indicate relatively higher productivity in comparison to the rest of the structures. 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 is the ultimate "stable point of attraction" for all the other structures. Note that, the concentration of 4-cliques is not very high (see Fig. 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 at this highly productive penultimate configuration. Another important point that is reflected in the year-wise analysis is that while 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 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. (2011) where they attribute that the box motif has maximum productivity as compared to others.

5.3 Longevity

The goodness of a group collaboration can also be captured by its longevity, i.e., the time the group has sustained without any structural imbalance. This point has been addressed by the social scientists multiple times by analyzing the longevity as a factor of the off-line groups to attract new members (Kairam et al. 2012). The proliferation of citation network and groups, however, has created new opportunities to study, at a large scale and with very fine resolution, the mechanisms which lead to characterize a

successful collaboration in a group level. 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 a 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 $\mathrm{Cr}(e_1)$, $\mathrm{Cr}(e_2)$ and $\mathrm{Cr}(e_3)$ [say, $\mathrm{Cr}(e_2) \geq \mathrm{Cr}(e_1), \mathrm{Cr}(e_3)$], respectively, and the times when two end researchers of an edge published their last paper together with these three collaborations by $\mathrm{Dl}(e_1)$, $\mathrm{Dl}(e_2)$ and $\mathrm{Dl}(e_3)$ [say, $\mathrm{Dl}(e_1) \leq \mathrm{Dl}(e_2), \mathrm{Dl}(e_3)$], respectively. Then the longevity of M is $(\mathrm{Dl}(e_1) - \mathrm{Cr}(e_2)) + 1$ (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. Formally, the *longevity* (τ) of a motif M is defined by the following equation:

$$\tau(M) = \min(\mathrm{Dl}(e_i)) - \max(\mathrm{Cr}(e_i)) + 1, \quad \forall e_i \in M, \quad (3)$$

where $\mathrm{Cr}(e_i)$ and $\mathrm{Dl}(e_i)$ denote the creation and 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 $\mathrm{Cr}(e_1) = 1972$, $\mathrm{Cr}(e_2) = 1973$, $\mathrm{Cr}(e_3) = 1974$, $\mathrm{Dl}(e_1) = 1976$, $\mathrm{Dl}(e_2) = 1979$ and $\mathrm{Dl}(e_3) = 1984$, then according to Eq. (3), 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 Sect. 4, we completely ignore such motifs in all our experiments.

5.4 Distribution of longevity

We plot the distribution of longevity for all motifs in Fig. 6. In each plot, we draw a vertical line to indicate the threshold that marks a long-lasting motif (mostly concentrated in the tail of the distribution) from the rest of the lot. Here also, the threshold is selected based on the value in the x-axis corresponding to which the first dip is observed in the line for most of the motifs (here the threshold is 5). Furthermore, we observe that the average longevity is highest for the 4-semi clique (6.72 years), followed by the 3-loop out motifs (6.63 years). Note that this result is quite unintuitive as one would, in general, anticipate that the 3-loop out and the 4-semi clique structures are incomplete in their construction as compared to the 3-clique and 4-clique structures, and should hence be less stable. Then to understand the average longevity of motifs over the years, we plot the year-wise longevity of each motif in Fig. 7. Here also, the results comply with the initial average behavior that the first three positions are dominated by semi-clique (Motif 7), 3-loop out (Motif 5) and box motif (Motif 6) in terms of the average longevity. Note that, all these three structures are very close to their penultimate stable form, i.e., the 4-clique. It seems that it becomes increasingly difficult to build new collaborations as a motif becomes more dense—this is similar to the idea of metastability that features quite often in real systems; the system becomes more rigid to change as it grows over time resulting in large relaxation times. The semi-clique has one edge left to form and therefore stays for the longest time in this state before it metamorphoses into the 4-clique through the formation of the last collaboration edge. The other two

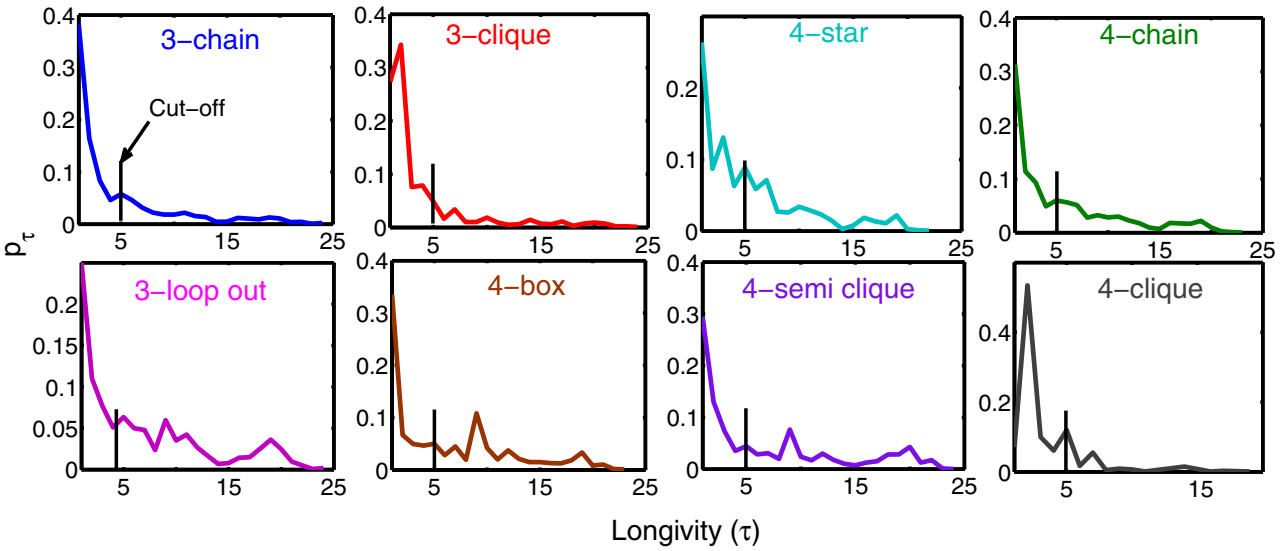


Fig. 6 (Color online) Longevity distribution of all motifs. The *horizontal line* in each frame indicates the cutoff based on which we develop the binary SVM classification model to predict the longevity of motifs

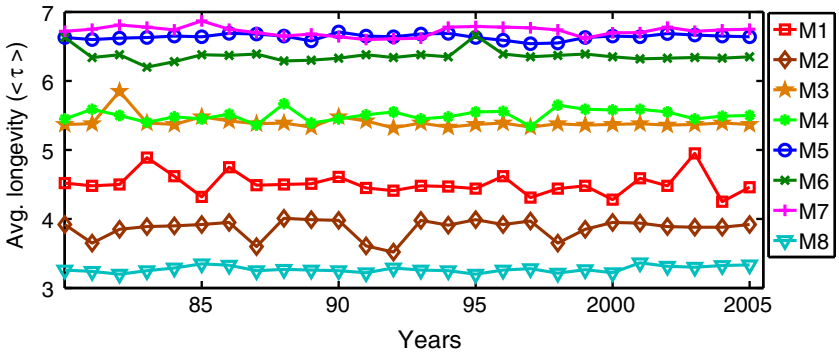


Fig. 7 (Color online) The average longevity per motif over the years (*Mi* stands for motif *i*)

structures are two edges away from the final configuration and stays longer than any other structure, but the semi-clique.

6 Author-level analysis

In this section, we address the first among the two principal objectives outlined in the introduction. In particular, we investigate the behavior of the individual authors within their local neighborhood. We observe that while the highly cited authors tend to be a part of a few specific motifs, namely, the 4-star, 4-clique and 3-loop out motifs, the motif distribution of the less-cited authors is almost uniform. It indicates a latent relationship between highly cited authors and high-productive motifs (star/4-clique motifs). Therefore, based on only the motif footprints, we try to design a supervised model that can efficiently predict the citation-based classification of authors. In this section, first we introduce several standard measures by means of which we evaluate the results obtained from all of our proposed predictive models and then we elaborate the model and the outcomes.

6.1 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:

$$\text{OA} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}. \quad (4)$$

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 toward a single class. For instance, if 95 % of the population belongs to the positive class (frequent class) and rest in the negative class (rare class),[4] randomly predicting all the samples as positive can produce a very high accuracy. In this case, the more challenging task is to precisely detect the negative classes. Therefore, to measure the performance of the system at a more granular level, we also estimate the following metrics along with the OA:

$$\text{Sensitivity}(R^+) = \frac{\text{Correctly classifed positive samples}}{\text{True positive samples}}, \quad (5)$$

$$\text{Specificity}(R^-) = \frac{\text{Correctly classifed negative samples}}{\text{True neagtive samples}}, \quad (6)$$

$$\text{PositivePrediction}(P^+) = \frac{\text{Correctly classified positive samples}}{\text{Positive classified samples}}, \quad (7)$$

$$\text{NegativePrediction}(P^-) = \frac{\text{Correctly classified negative samples}}{\text{Negative classified samples}}. \quad (8)$$

6.2 Author classification model

We build a supervised model using only the clues of the motif distribution that can classify the authors based on the number of citations they obtain. From the citation distribution, we empirically set up a threshold on the number of citations (here, we consider the citation threshold as 3500). If an author gets the total number citations more than the selected threshold, she is marked as a **highly cited author (negative class)**, otherwise she is marked as a **less-cited author (positive class)**. Naturally, the entire population is biased toward the positive class (88.44 % of the entire population). We use support vector machine (SVM) (Cortes and Vapnik 1995) as a supervised model to classify the authors. 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

[4] Note that, we refer to the rare class as negative class and frequently observed class as positive class in the rest of the paper.

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

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

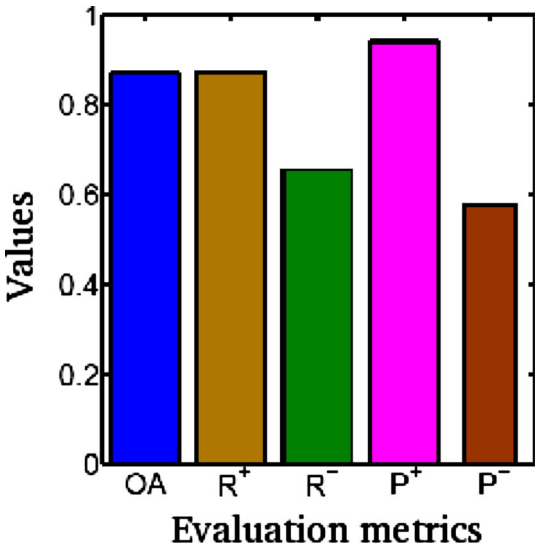


Fig. 8 (Color online) Performance analysis of author classification model (*Mi* stands for motif *i*)

kernel. In the training sample, each instance corresponds to an author, and the eight features correspond to the count of the eight different motifs to which the author belongs to. We adopt a tenfold cross-validation technique where the whole population is randomly partitioned into ten equal size subsamples. A single subsample is retained as the validation data for testing the model, and the remaining nine subsamples are used as training data. The cross validation process is then repeated ten times, with each of the ten subsamples used exactly once as the validation data. Then we average the ten results from the folds and plot in Fig. 8. We observe that while the overall accuracy of the model is 0.87, the model quite accurately identifies the instances of the rare class, i.e., the negative class ($R^- = 0.8732$, $P^- = 0.94$). Note that, such a high level of accurate classification is not observable through direct citation-based analysis. However, the current result immediately unfolds the fact that simple motif counts have a remarkable discriminative power that not only allows the model to make accurate predictions for the frequently observed class, but also for the rare class.

7 Motif characteristics

Toward the second objective outlined in the introduction, we first identify a set of discriminative features that could be attributed as characteristics of a group (i.e., a network motif). In this section, therefore, we analyze such a set of distinctive features of group collaborations derived from the characteristics of the constituent authors. We also analyze the correlation of the following features with two goodness metrics discussed in Sect. 5 for all the motifs.

7.1 Construction time (CT)

Since each individual edge in a motif indicates a one-to-one collaboration, each edge 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: $\mathrm{CT}(M) = \mathrm{Max}(\mathrm{Cr}(e_i)) - \mathrm{Min}(\mathrm{Cr}(e_i)) + 1, \forall e_i \in M$, where $\mathrm{Cr}(e_i) =$ year of creation of edge $e_i \in M$. For example, if a 3-node motif M is constructed by three edges e_1, e_2 and e_3, and $\mathrm{Cr}(e_1) = 1972$, $\mathrm{Cr}(e_2) = 1973$, $\mathrm{Cr}(e_3) = 1974$, then the construction time of M is $\mathrm{CT}(M) = (1974 - 1972) + 1 = 3$ years.

We intend to examine whether the construction time has any effect on the productivity and longevity of a motif. The top two frames in the first column of Fig. 9 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. However, the longevity distribution versus construction time in the bottom frame of the first column in Fig. 9 depicts that for all the motifs, the longevity increases with the increase in construction time. In other words, a motif that has taken a larger time to come into existence usually persists for a larger time in future. The possible reason could be that a longer construction time allows the constituent authors to build trust among each other resulting in a large persistence of the collaboration in future.

7.2 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 is the motif indicating 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 both productivity and longevity. The top two frames in the second column of

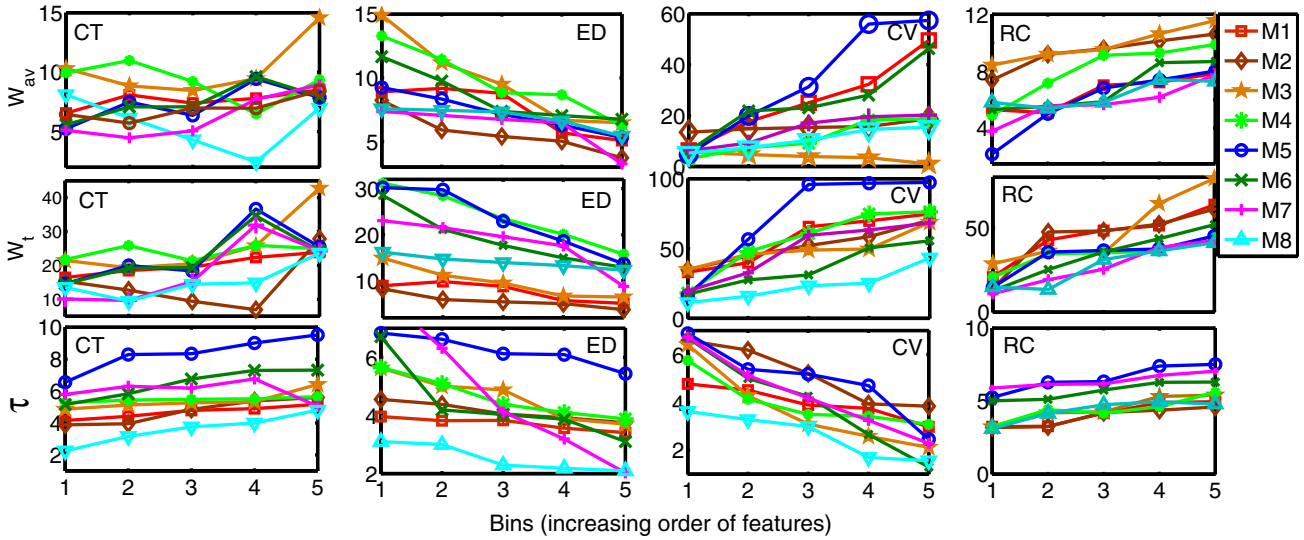


Fig. 9 (Color online) Variation of productivity (both W_{av} and W_t) and longevity of motifs with different characteristic features mentioned within each frame. All motif instances are grouped into five bins (1:low, 5:high) according to the values of the corresponding characteristic features (*Mi* stands for motif *i*)

Fig. 9 show the productivity of all the motifs arranged in various ranges of experience diversity. We observe that the average productivity (for both the measures) decreases with the increase of experience diversity of a group. A similar pattern is observed in the bottom frame of the second column of Fig. 9 where the longevity also declines with increasing degree of experience variance. From these, we might conclude that the groups comprising peer researchers of similar experience are much more productive and the collaborations are sustained for a longer period of time compared to the groups led by a single experienced researcher. The possible reason could be that groups of researchers with the same research experience usually have enough scope of open discussion and arguments that would eventually lead to productive outcomes and to retain the groups intact for a long time; on the other hand, this is mostly missing in a highly diverse group of people where the group is mostly dominated by the most experienced researchers.

7.3 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 a 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 and longevity of a group collaboration, i.e., are the groups containing all highly cited researchers superior than the less-cited groups? The top two frames of the third column in Fig. 9 show that except in star motif (Motif 3), all other motifs show a consistent pattern that average productivity increases with the increase of citation variance. This result is markedly in contrast to the earlier results described in Sect.7.2. 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 detail in the feature correlation subsection of Sect. 8. However, the longevity declines with the increase in citation variance (bottom frame in the third column of Fig. 9) which is similar to the earlier result of experience diversity.

7.4 Recency (RC)

As the citation counts accumulate over time, it is important to have a measure of the age of a group and to observe its relationship with the longevity and the productivity metrics. 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 of edges]. We study as a feature the number of years since the motif was fully created. 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. We expect that the longer a group (motif) stays, the more

citations it might receive. The top two frames of the last column in Fig. 9 show that the productivity of all the motifs increases with the increase in the stabilization time. The possible reason could be that as long as a motif is stable and does not get converted to other motifs, it keeps on producing effective results. However, no such strong correlation is observed between this feature and the longevity of a collaboration; in fact, longevity seems to remain flat as one varies RC (bottom frame of the last column of Fig. 9).

7.5 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 convert into another. We have already mentioned the use of motif transition in Sect. 7.4 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 into 3-clique. For 4-node motifs, the process follows a little complicated dynamics as shown in Fig. 10 (upper). For instance, addition of *single edge* in the system one at a time can lead to any of the following three paths (or the sub-paths): $a-c-e-f$, $b-d-e-f$ and $b-c-e-f$. However, in the real-world scenario, it can be possible that more than one edge gets 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 Fig. 10 (lower). Here, if n nodes in a motif M at time t_i get 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 the 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 11a 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

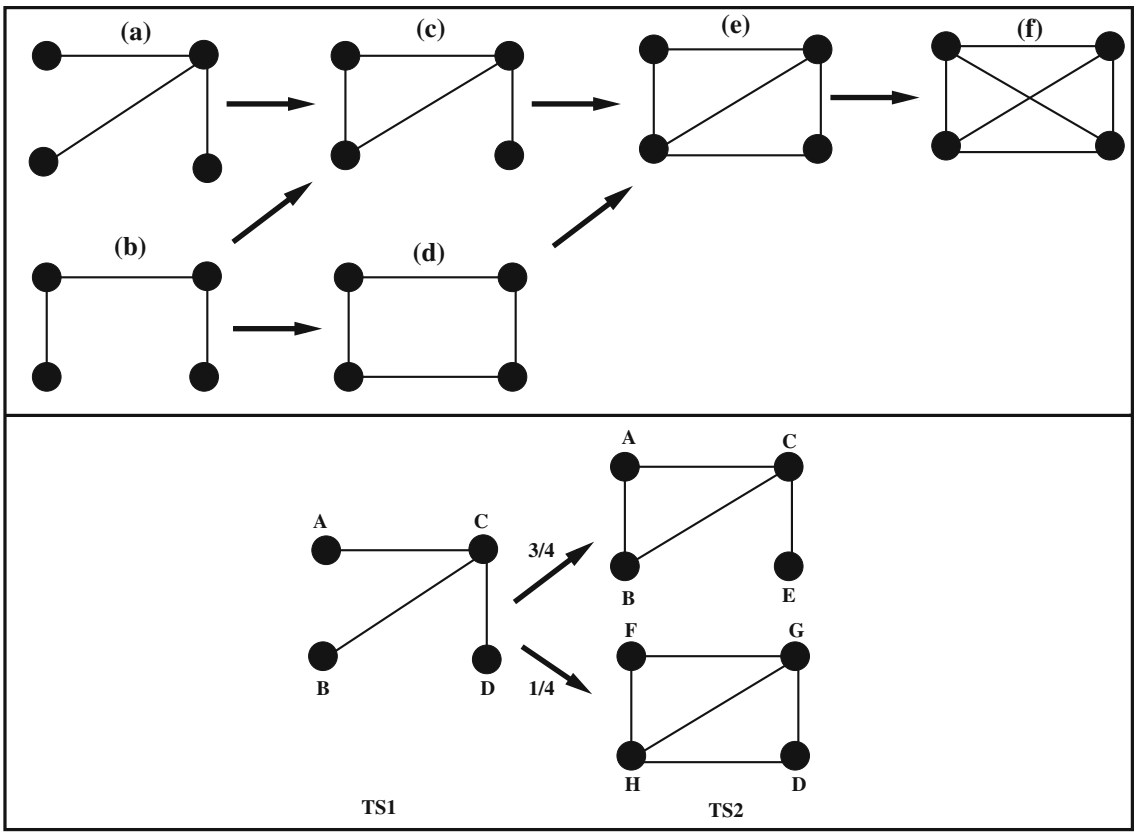


Fig. 10 Transition of 4-node motifs after adding a single edge (*upper*) and the toy example of mapping motif(s) across successive timestamps (TS) (*lower*)

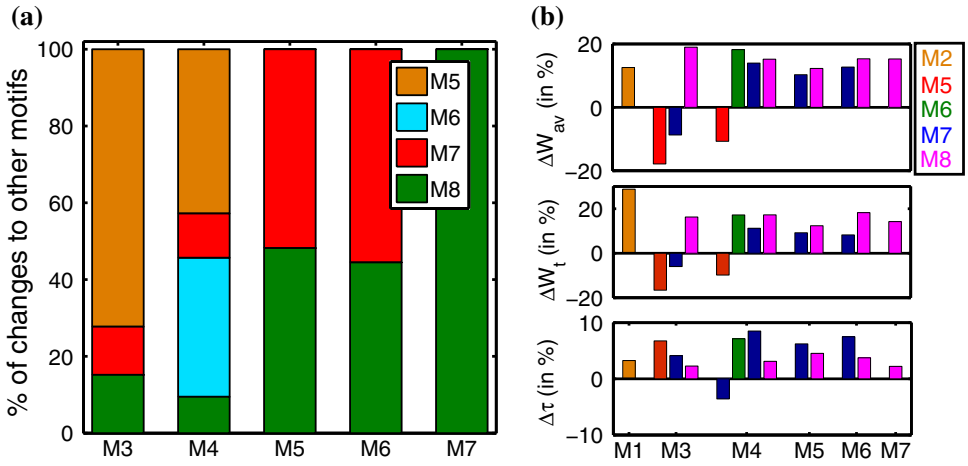


Fig. 11 (Color online) **a** Fraction (in %) of changes of one motif into the others across all time transitions and **b** gain in productivity and longevity due to motif transition (*Mi* stands for motif *i*)

"add-edge one" behavior discussed in Fig. 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. The possible reason is that once a researcher chooses her immediate collaborators, she would become a good proxy for those collaborators to build new collaborations among themselves. However, new collaborations are usually formed one at a time. For instance, when a 4-star motif is created, the possibility of one edge to be added among the neighbors around the center node tends to become higher, which results in the formation of 3-loop out motif. Therefore, 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: $\Delta W = \frac{W_{\text{new}} - W_{\text{old}}}{W_{\text{old}}}$ (W can be replaced by W_{av} or W_t). Similarly, we measure the gain/lose of longevity due to motif transition. Figure 11b 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 toward this structure. On the other hand, the average time of longevity increases for most of the cases due to the motif transition. Some interesting observations here are that the productivity increases when a star motif gets converted to a clique motif, although in general astar motif is more productive than a clique motif (see Fig. 5). However, the chance of this conversion is rare (see Fig. 11); hence in most cases, the clique motif appears after passing through several other intermediate motif configurations with subsequent decrease in the productivity.

8 Group-level analysis

In this section, we discuss two predictive models that can help forecast the longevity and productivity of a motif by analyzing a set of discriminating features discussed in Sect. 7. Essentially, for a single motif, the following features are used in these models: construction time (CT), experience diversity (ED), citation variance (CV), recency (RC), average productivity (W_{av}), change in average productivity (ΔW_{av}), total productivity (W_t), change in total productivity (ΔW_t), longevity (τ) and change in longevity ($\Delta\tau$), while longevity and $\Delta\tau$ are only used in productivity prediction model, the four features related to productivity (W_{av}, W_t, ΔW_{av} and ΔW_t) are used only in the longevity prediction model. the rest of the features are used in both models. It is important to note that while computing ΔW_{av} and ΔW_t of a motif, instead of considering to which motif(s) the present motif would transform at a later time, we consider the previous history of the motif, i.e., from where the motif was itself created. We adopt this policy so that we could refrain from using quantities that are observable only in future time points, since it is inappropriate (leading to information leakage) to employ quantities from the future to predict the future. Note that the motivation behind developing these two prediction models is as follows: if we know the longevity and the productivity of different types of collaborations (represented by motifs), it would be helpful for new researchers to gain ideas on (i) how to build effective collaborations, (ii) which collaborations they should maintain, (iii) what is the dynamics that could lead to an effective collaborations, etc.

8.1 Feature correlations

Before entering into the detailed description of the two models, we perform a systematic analysis of the correlations between the features to identify if any of the features

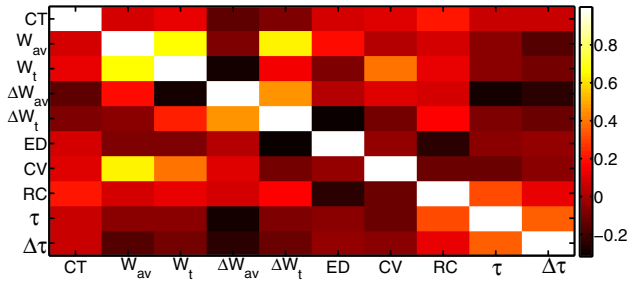


Fig. 12 (Color online) Heat map depicting the correlations among different features

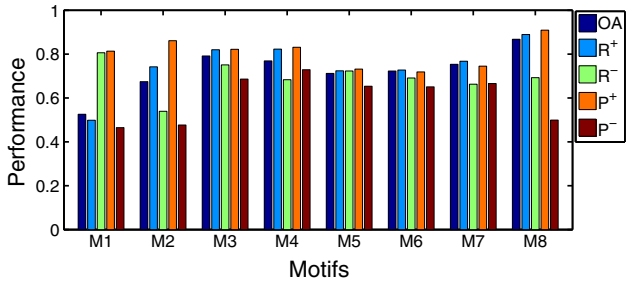


Fig. 13 (Color online) Performance of the SVM model to predict the longevity of motifs (*Mi* stands for motif *i*)

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 Fig. 12. We observe the maximum correlation between W_{av} and W_t (0.67), followed by ΔW_{av} and ΔW_t (0.293) which is also quite intuitive, since both of these are derived from the same hypothesis. The highest negatively correlated pair is ΔW_t and ED (-0.32), followed by ΔW_{av} and τ (-0.29). Most of the correlations among the pairs of features are very small or negative, which implies that the feature set is highly discriminative and uncorrelated. 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 with some of them in lieu of the other. Therefore, we use all the features in the subsequent analysis and predictions made in the rest of this section.

8.2 Model 1: longevity prediction model

We develop a supervised binary classification model to predict the longevity of a motif. To decide the cutoff among the spectrum of longevity values of motifs shown in Fig. 6, we observe that in most of the cases, the first dipping of the distribution of longevity occurs at the value of 5 in the x-axis. Therefore, we consider all motifs having $\tau < 5$ as "short-lived motifs" (frequent class, positive class, short-term collaborations) and others as "long-lived motifs" (rare class, negative class, long-term collaborations). From Fig. 6, it is apparent that the population is highly biased toward the positive class. Here, we retain our earlier experimental setup discussed in Sect. 6. The performance of the classifier after tenfold cross validation is measured for each of the motifs separately and pictorially depicted in Fig. 13. We observe that while the average overall accuracy of the system is 0.72, the system performs reasonably well to predict the longevity of 4-cliques (OA $= 0.87$, $R^+ = 0.89$, $R^- = 0.69$, $P^+ = 0.91$ and $P^- = 0.50$). For most of the motifs, the sensitivity (R^+) and PositivePrediction (P^+) of the model are above 0.70. 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 highest 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 the overall performance of the model for 4-node motifs is reasonably high compared to the 3-node motifs.

Error analysis We systematically analyze the significance of the features used in this model by dropping them one at a time and measuring the performance of the model. In Fig. 14, we plot the average error (with standard deviation) that occurs due to the drop of each of the features. Each frame in the figure corresponds to the error due to the drop of one feature mentioned in the frame. For better comparison, in each frame we also plot the average error which occurs when all the features are used (broken line). From the error analysis, it is apparent that the features related to productivity (W_{av} and W_t) significantly contribute in predicting the accurate results. Surprisingly, in certain cases (for Motif 4), dropping the construction time from the feature set enhances the performance of the system. We have mentioned earlier in Fig. 9 (Sect. 7) that the construction time does not indicate any uniform pattern within or across different classes of motifs. The results of error analysis also corroborate this observation, thus pointing to the fact that this feature does not have strong discriminative power.

8.3 Model 2: productivity prediction model

The second model is again a binary classifier that tries to classify the motifs based on their productivity. Here also, we detect the threshold for binary classification of productivity similarly as in the earlier experiment. We observe that both W_{av} and W_t follow similar distributions. In Fig. 4, we plot only the distribution of W_{av} to decide the threshold. The threshold is decided to be 10, i.e., the motifs having $W_{\text{av}} < 10$ are considered as "low productive" (positive

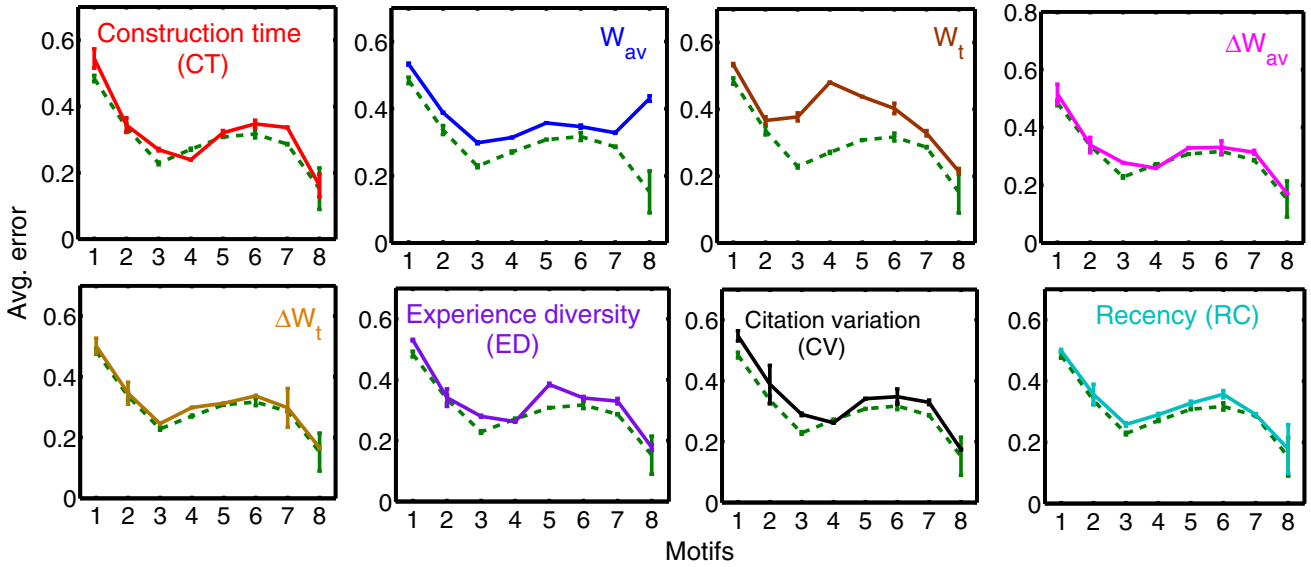


Fig. 14 (Color online) Error analysis of the longevity prediction model. *Each frame* shows the average error (with *vertical bars* representing standard deviation of errors) of the model when removing the corresponding feature mentioned in the frame. The *broken green line* depicts the average error when using all features. The number i in the x-axis stands for Motif i

class) and the rest as "high productive" (negative class). Again, the system is mostly biased toward the positive class. Here also, we use SVM with linear kernel and the results are reported after tenfold cross validation. Note that, in this model we use six features, namely construction time, experience diversity, citation variance, recency, longevity and change in longevity to predict the productivity (W_{av}). Figure 15 presents the accuracy of the model for each of the motifs. On an average, the performance of the second model is better than the earlier model where most of the values cross 80 % accuracy. Here also, the model more accurately predicts the productivity of 4-cliques (OA $= 0.95$, $R^+ = 0.98$, $R^- = 0.62$, $P^+ = 0.96$ and $P^- = 0.74$). While R^- is greater than 60 % throughout, P^- for the 4-node motifs is greater than 60 %.

Error analysis Since the performance of the second model is superior to the earlier one in spite of the less number of features used, it would be interesting to analyze the influence of each feature to enhance the performance of the prediction model. For this, we again measure the importance of each feature by dropping one at a time and comparing the relative decline of the average accuracy. Figure 16 displays the error that occurs due to omitting each feature. Here, while four features, namely construction time, experience diversity, recency and longevity, show their reasonable importance in predicting the productivity, the citation variance proves to be immensely important in this model for all the motifs. More particularly, for star motif (Motif 4), dropping the citation variance can degrade the performance of the model nearly three times lower than the original. This result not only

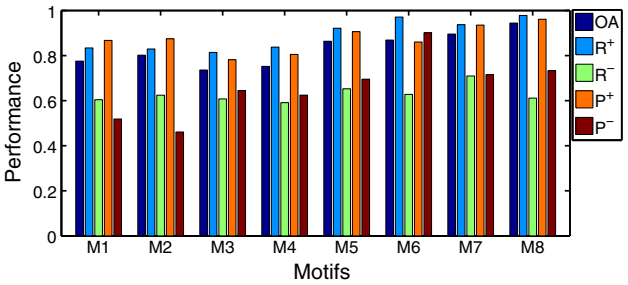


Fig. 15 (Color online) Performance of the SVM model to predict the productivity of motifs (*Mi* stands for motif i)

signifies the extent of importance of this feature among the others, but also reflects its enormous power of predicting productivity of group collaborations.

9 Correlating longevity and productivity

In this section, we finally look back at the entire population of motifs and try to relate the two goodness measures of collaborations. Since we classified the entire dataset into two classes in terms of longevity and productivity individually, the entire population can be divided further into four regions: long-term high productive collaborations, long-term low productive collaborations, short-term high productive collaborations and short-term low productive collaborations. In Table 3, we present two confusion matrices for 3-node and 4-node motifs separately showing the fraction of population in each of the regions obtained directly from the real dataset (i.e., gold-standard). Thus, the interpretation of an individual

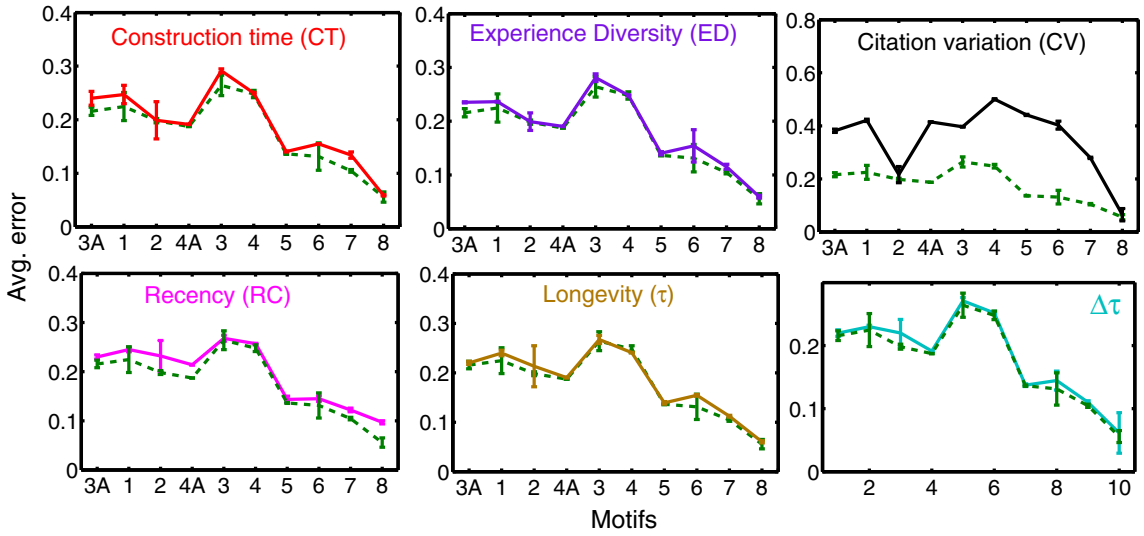


Fig. 16 (Color online) Error analysis of the productivity prediction model. *Each frame* shows the average error (with *vertical bars* representing the standard deviation of errors) of the model when removing the corresponding feature mentioned in the frame. The broken *green line* depicts the average error when using all the features (*Mi* in *x*-axis stands for motif *i*)

Table 3 Confusion matrices showing four regions of the population of the gold-standard dataset

Longevity	Productivity	
	Low	High
(a) 3-node motifs		
Short	0.59	0.15
	(0.54)	(0.12)
Long	0.20	0.06
	(0.25)	(0.09)
(b) 4-node motifs		
Short	0.44	0.16
	(0.44)	(0.17)
Long	0.29	0.11
	(0.32)	(0.07)

The values within parenthesis are obtained from the two prediction models

entry in this matrix for the 3-node (4-node) motifs is as follows—59 % (44 %) of the motifs are short lived as well as less productive, 15 % (16 %) are short lived but highly productive, 20 % (29 %) are long lived but less productive and only 6 % (11 %) are long lived as well as highly productive. The values in Table 3 within parenthesis are the same quantities obtained from our prediction models. As discussed earlier, for both the models, we have used tenfold cross validation on the entire dataset and performed 50 such iterations to obtain the average performance. Therefore, in each iteration, one-tenth of the entire population serves as the test dataset and we obtain 50 such confusion matrices. The values within parenthesis in Table 3 are the average of these confusion matrices. It is evident from the table that the results obtained from the models are remarkably similar to those indicated by the gold-standard statistics. From these tables, we conclude with reasonably high confidence that short-term collaborations are generally less productive. One general argument could be that since short-term collaborations generally do not persist for long, the scope of the growth of productivity is severely limited. This immediately points to the fact that short-term collaborations would generally not lead to the production of high-quality research output leading to very low gain in citations. Further, there is a feedback effect in that if a collaboration is repeatedly failing to produce high-quality output, the chances that it would persist longer automatically diminishes. On the other hand, we cannot make any strong conclusion regarding long-term collaborations due to the lack of enough statistical evidences. In short, the main observation here is that none of the goodness metrics can alone completely determine the other.

10 Motif distributions for different fields

The rich metadata information of our dataset further allows us to measure the distribution of motifs for different fields of computer science domain. Since we know the field of research for a particular paper present in our dataset, we mark each author by the field in which she has published maximum papers. Following this, a motif is marked by the field which is the research interest of majority of its constituent authors. In case there is a tie, we resolve by marking the motif as a part of all the fields that are in tie. Therefore, each field now constitutes motifs of different kinds.

Table 4 Number of motifs of different kinds in each fields of computer science domain

Fields	Motif 1	Motif 2	Motif 3	Motif 4	Motif 5	Motif 6	Motif 7	Motif 8
SC	1,688,886	194,877	53,383,245	52,474,614	11,776,366	23,359	454,191	289,657
IR	868,678	62,076	37,947,595	39,805,361	6,923,663	25,124	269,414	94,430
WWW	309,723	30,174	14,290,517	15,869,177	3,591,709	15,370	167,841	51,526
SEC	1,335,591	74,987	49,576,069	60,556,754	8,665,075	51,306	323,305	89,930
EDU	843,551	92,844	18,846,701	25,900,113	5,866,133	18,012	210,945	123,778
DIST	4,848,263	288,864	2,50,298,270	2,47,938,091	40,606,341	133,592	1,527,534	497,202
PL	649,086	48,650	17,652,040	26,929,675	4,408,607	20,600	153,060	69,716
ALGO	4,470,929	220,619	1,54,086,899	2,09,587,945	25,720,651	256,572	990,753	262,807
NETW	12,751,861	578,930	6,38,477,904	6,52,386,924	73,236,665	298,477	2,128,468	699,272
ML	2,534,953	1,80,008	1,08,080,837	1,01,363,927	15,556,938	44,520	562,106	225,303
MUL	2,358,571	236,411	1,09,790,561	1,12,837,986	18,697,906	65,554	808,017	407,547
DB	4,904,153	243,998	2,36,692,742	2,74,019,047	36,943,158	202,607	1,230,156	349,173
HCI	2,853,505	191,620	1,11,550,010	1,15,121,687	19,951,664	68,071	710,443	270,388
NLP	4,205,876	289,228	1,29,624,546	1,57,391,508	26,957,671	100,780	1,031,139	392,589
AI	12,694,180	669,806	5,61,295,885	5,58,603,229	73,590,521	248,410	2,500,253	875,519
EMB	783,880	38,291	26,096,515	28,746,135	4,128,207	14,159	141,237	38,218
BIO	2,118,607	328,327	62,475,465	68,578,080	17,897,512	19,538	608,051	502,223
ARCH	6,823,624	435,287	2,80,054,454	2,63,184,366	42,823,227	117,702	1,512,624	589,613
DM	902,885	72,630	59,850,936	55,049,523	7,691,422	26,924	227,344	99,499
GRPH	2,217,349	121,457	91,395,885	93,698,746	15,010,651	51,757	530,306	160,896
SE	3,915,726	247,149	1,35,119,019	1,61,843,896	24,726,027	90,184	852,323	3,08,106
CV	1,776,528	94,266	8,8,906,676	83,042,560	12,292,303	41,360	413,241	127,687
SIM	308,241	29,598	7,987,495	9,930,622	1,911,701	3833	64,057	38,394
OS	242,635	24,192	8,130,245	12,016,895	2,399,418	8987	93,908	42,784

Table 4 shows the count of motifs of different kinds in each field of the computer science domain. We also plot the fraction of each 3-node motif among all 3-node motifs and fraction of each 4-node motifs among all 4-node motifs (the global pattern is shown in Fig. 3). We observe two distinct patterns as shown in Fig. 17. For few fields such as information retrieval, World Wide Web etc. (as denoted by red colored line), the proportion of 4-chain motif is higher than that of 4-star motif, whereas for few fields such as scientific computing, networking, etc. (as denoted by green colored line), the proportion is just opposite. Since the reason is not very clear to us, more deeper analysis on the motifs in different fields would remain as a potential area to be unfolded in the future.

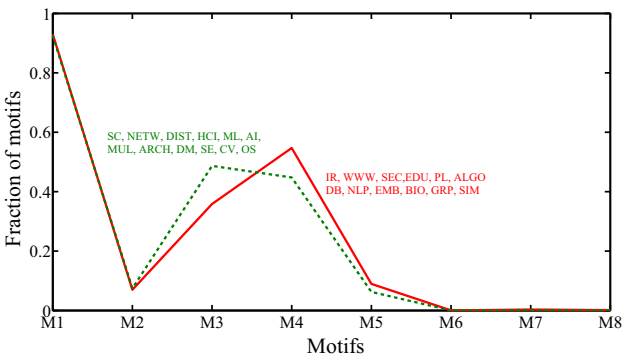


Fig. 17 (Color online) Fraction of motifs in different fields of the computer science domain. The fraction is calculated independently for 3-node and 4-node motifs

11 Real examples 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 and 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

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

observe another star motif constituting A. Aggarwal, M. Charikar and D. Williamson in the 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) to gain productivity.

- *Jiawei Han's collaborations* We have noticed in Sect. 7 that a 4-chain tends to get converted to a 4-clique motif more often to gain high productivity. A prominent evidence found in our dataset is the 4-chain comprising Jiawei Han,[8] Yongjian Fu, Zhaohui Xie and Wei Wang (where three collaboration edges are formed between first–second, second–third and third–fourth 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".
- *James Allan's collaborations* Similar to the earlier observations, James Allan[10] is found to occur a maximum number of times in 4-cliques. However, the 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 significantly higher productivity. This is counterintuitive to our earlier observation in Sect. 7 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.

12 Extending the results to physics dataset

To check the robustness of the important conclusions drawn from Sects. 5 to 8, we conduct a shallow analysis on Physics dataset. We use all published articles in Physical Review (PR) journals[12] from 1975 till the end of 2010. We use all such entries which possess the information about their index, title, name of the author(s), year of publication and references. The filtered dataset contains 325,399 valid papers and 277,154 authors. After author–name disambiguation, we construct collaboration network and extract all 3-node and 4-node motifs separately. The fraction of motifs in each category is reported in Table 5.

Here, we verify few important results that were previously observed in the computer science dataset.

- Here also, we observe that the average productivity is higher for star motif ($W_{\text{av}} = 10.28, W_t = 22.45$), which is followed by 4-clique ($W_{\text{av}} = 9.11, W_t = 20.15$) and 4-chain motifs ($W_{\text{av}} = 8.97, W_t = 18.44$). 4-box motif seems to be least productive ($W_{\text{av}} = 4.11, W_t = 12.31$).
- In terms of average longevity, 4-semi clique and 3 loop out motifs seem to be highest (6.82 and 6.21 respectively), where the 4-clique seems to diminish quickly (longevity of 2.15).
- We further build the author classification model keeping the citation threshold as 3500 to separate highly cited and low cited authors, and run tenfold cross validation. It turns out to be very effective in terms of standard

Table 5 Percentage of motifs in the Physics collaboration network

Motifs	Types	Percentage
3-node motifs	3-chain	65.28
	3-clique	34.72
4-node motifs	4-star	25.47
	4-chain	28.91
	3-loop out	26.10
	4-box	4.48
	4-semi clique	6.28
	4-clique	8.76

[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/.

[12] http://journals.aps.org/datasets.

Table 6 Matrix showing the transition of motifs

	M3	M4	M5	M6	M7	M8
M3	–	–	**82.91**	–	4.47	12.22
M4	–	–	34.87	**38.98**	10.23	15.92
M5	–	–	–	–	**76.54**	23.46
M6	–	–	–	–	**82.13**	17.87
M7	–	–	–	–	–	**100**

The row corresponds to the initial form of the motif and the column corresponds to the final form of the motif. The maximum percentage of transition for a particular motif is highlighted in bold font

evaluation metrics ($\text{OA} = 0.85, R^+ = 0.79, R^- = 0.60, P^+ = 0.91, P^- = 0.56$) to classify authors based on the surrounding motif distribution.

- We examine each of the proposed features such as construction time, experience diversity, citation variance and recency, and find the correlation with the productivity and longevity. Except a couple of cases such as experience diversity and citation variance where the evidence is less prominent, for other cases the correlation highly corroborates with the results observed earlier for the computer science dataset. For instance, longevity and productivity seem to increase with the increase of construction time and recency.
- Regarding the motif transition, we earlier observed that motif evolution is not abrupt, but follows typical "add-edge one" mechanism. Here surprisingly, the previous observation remains persistent with high statistical significance. A broad experimental result is presented in Table 6. For motifs M5 and M6, in around 80 % cases these two motifs transform to M7, which is much stronger evidence compared to the earlier observation in computer science dataset where the chance was nearly 50 %. The reason could be that physics is a much older field than computer science, thus facilitating this phenomenon for a longer time.
- We further examine the gain/loss in productivity and longevity due to motif transition. We observe nearly 14.45 and 10.43 % average gain in productivity in terms of W_{av} and W_t, respectively, while a motif gets transformed into 4-clique. On the other hand, maximum gain in longevity (8.74 %) is observed when a 4-box motif gets converted into 4-semi clique.
- Finally, we run two group-level models for predicting longevity and productivity of different motifs. We keep the thresholds mentioned in Sects. 8.2 and 8.3 for dividing the population into two classes. The average accuracy after tenfold cross validation is reported in Table 7. We observe that the performance of both the models is significantly well in terms of all the validation measures.

Table 7 Accuracy of (left) longevity and (right) productivity prediction models of motifs on the physics dataset

Motifs	OA	R^+	R^-	P^+	P^-
Longevity prediction model					
M1	0.60	0.58	0.72	0.71	0.75
M2	0.82	0.85	0.75	0.89	0.82
M3	0.85	0.89	0.91	0.90	0.88
M4	0.78	0.82	0.89	0.79	0.80
M5	0.81	0.76	0.78	0.81	0.79
M6	0.75	0.81	0.71	0.83	0.87
M7	0.82	0.85	0.76	0.79	0.82
M8	0.89	0.87	0.82	0.85	0.81
Productivity prediction model					
M1	0.81	0.89	0.71	0.80	0.78
M2	0.85	0.82	0.69	0.81	0.61
M3	0.70	0.79	0.65	0.80	0.65
M4	0.75	0.81	0.63	0.78	0.69
M5	0.86	0.82	0.79	0.81	0.63
M6	0.82	0.88	0.85	0.83	0.79
M7	0.88	0.89	0.74	0.81	0.78
M8	0.92	0.91	0.87	0.82	0.89

The above observations indicate that most of the conclusions drawn in this paper are highly robust and applicable for different domains.

13 Conclusions

In this work, we showed that in the collaboration network, the network contexts of individuals represented by network motifs have significant potential to unfurl the underlying dynamical behavior of authors within a group and along with the group as a whole. We further established that it is indeed possible to go beyond pairwise collaborations and investigate the fundamentals of various types of group collaborations represented in the form of motifs. We conclude the paper mentioning few interesting outcomes and some immediate future directions as follows: (i) we observe that while star and the 4-clique motifs are highly productive, semi-clique motifs seem to have a very high longevity, (ii) the productivity of a motif is not random; rather it is driven by the structural and functional importance of the different collaborations, (iii) the distribution of network motifs neatly classifies the highly cited authors from the rest, (iv) the characteristics of a group collaboration can suitably be explained in terms of a set of distinctive features of the constituent researchers, (v) in real world, the transition of motifs over the successive time steps is usually not abrupt, rather it systematically follows "add-edge one" mechanism, (vi) transition to a 4-clique

produces the largest gain in productivity for all the motifs, (vii) the characteristic features of the motifs quite efficiently predict the longevity and productivity of a group collaboration with the best predictions being for the 4-clique.

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 (Krumov et al. 2011). 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

Abbasi A, Chung KSK, Hossain L (2012) Egocentric analysis of co-authorship network structure, position and performance. Inf Process Manag 48(4):671–679

Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet 8(6):450–461

Backstrom L, Leskovec J (2011) Supervised random walks: predicting and recommending links in social networks. In: WSDM. ACM, New York, NY, USA, pp 635–644

Baras JS, Hovareshti P (2011) Motif-based communication network formation for task specific collaboration in complex environments. In: ACC 2011. IEEE, Kerala, India

Biryukov M (2008) Co-author network analysis in dblp: classifying personal names. In: MCO. Springer, Berlin, pp 399–408. http://link.springer.com/chapter/10.1007%2F978-3-540-87477-5_43

Chakraborty T, Ganguly N, Mukherjee A (2014) Automatic classification of scientific groups as productive: an approach based on motif analysis. In: 2014 IEEE/ACM international conference on advances in social networks analysis and mining, ASONAM 2014, Beijing, China, August 17–20, 2014, pp 130–137

Chakraborty T, Sikdar S, Tammana V, Ganguly N, Mukherjee A (2013) Computer science fields as ground-truth communities: their impact, rise and fall. In: Advances in social networks analysis and mining 2013, ASONAM '13, Niagara, ON, Canada—August 25–29, 2013, pp 426–433

Chakraborty T, Tammana V, Ganguly N, Mukherjee A (2015) Understanding and modeling diverse scientific careers of researchers. J Informetr 9(1):69–78. doi:10.1016/j.joi.2014.11.008. http://www.sciencedirect.com/science/article/pii/S1751157714001102

Choobdar S, Ribeiro P, Bugla S, Silva F (2012) Comparison of co-authorship networks across scientific fields using motifs. In: ASONAM. IEEE Computer Society, Los Alamitos, pp 147–152

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297

Dascal M (1989) On the roles of context and literal meaning in understanding. Cogn Sci 13(2):253–257

Ding Y (2011) Scientific collaboration and endorsement: network analysis of coauthorship and citation networks. J Informetr 5(1):187–203

Hyun Yook S, Oltvai ZN, lszl Barabsi AL (2004) Functional and topological characterization of protein interaction networks. Proteomics 4:928–942

Han Y, Zhou B, Pei J, Jia Y (2009) Understanding importance of collaborations in co-authorship networks: a supportiveness analysis approach. In: SDM. Springer, Berlin, pp 1111–1122

Huang J, Zhuang Z, Li J, Giles CL (2008) Collaboration over time: characterizing and modeling network evolution. In: WSDM. ACM, New York, pp 107–116

Kairam SR, Wang DJ, Leskovec J (2012) The life and death of online groups: predicting group growth and longevity. In: Proceedings of the fifth ACM international conference on web search and data mining, WSDM '12. ACM, New York, NY, USA, pp 673–682. doi:10.1145/2124295.2124374

Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics 20(11):1746–1758

Kronegger L, Mali F, Ferligoj A, Doreian P (2012) Collaboration structures in slovenian scientific communities. Scientometrics 90(2):631–647

Krumov L, Fretter C, Müller-Hannemann M, Weihe K, Hütt M (2011) Motifs in co-authorship networks and their relation to the impact of scientific publications. EPJB 84(4):535–540

Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am Soc Inf Sci Technol 58(7):1019–1031

Liu J, Lei KH, Liu JY, Wang C, Han J (2013) Ranking-based name matching for author disambiguation in bibliographic data. In: Proceedings of the 2013 KDD cup 2013 workshop, KDD Cup '13. ACM, New York, NY, USA, pp 8:1–8:8. doi:10.1145/2517288.2517296

Liu HT, Pei D, Wu Y (2012) A novel evolution model of collaboration network based on scale-free network. ICHIT 2:148–155

Lü L, Zhou T (2010) Link prediction in weighted networks: the role of weak ties. EPL 89(1):18,001. http://stacks.iop.org/0295-5075/89/i=1/a=18001

Martinez-Romo J, Robles G, González-Barahona JM, Ortuño-Perez M (2008) Using social network analysis techniques to study collaboration between a floss community and a company. In: OSS. Springer, Berlin, pp 171–186

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827

Newman MEJ (2001) The structure of scientific collaboration networks. PNAS 98(2):404–409

Newman M (2004) Coauthorship networks and patterns of scientific collaboration. PNAS 101:5200–5205

Pan RK, Saramäki J (2011) The strength of strong ties in scientific collaboration networks. CoRR. abs/1106.5249

Prill RJ, Iglesias PA, Levchenko A (2005) Dynamic properties of network motifs contribute to biological network organization. PLoS Biol 3(11):e343

Rennie JDM, Srebro N (2005) Fast maximum margin matrix factorization for collaborative prediction. In: ICML. ACM, New York, pp 713–719

Hassan S-U, Ichise R (2009) Discovering research domains using distance matrix and co-authorship network. SDM 3:1252–1257

Said YH, Wegman EJ, Sharabati WK, Rigsby JT (2008) Social networks of author-coauthor relationships. Comput Stat Data Anal 52(4):2177–2184

Shi X, Wu L, Yang H (2008) Scientific collaboration network evolution model based on motif emerging. In: ICYCS. IEEE Computer Society, Washington, pp 2748–2752

Tambayong L (2007) Dynamics of network formation processes in the co-author model. J Artif Soc Soc Sim 10(3):2. http://dblp.uni-trier.de/db/journals/jasss/jasss10.html#Tambayong07

Wernicke S (2005) A faster algorithm for detecting network motifs. In: WABI. Springer, Berlin, pp 165–177

Wernicke S, Rasche F (2006) Fanmod: a tool for fast network motif detection. Bioinformatics 22(9):1152–1153

Wu, W., Han, Y., Li, D.: The topology and motif analysis of journal citation networks. In: CSSE, pp. 287–293. IEEE Computer Society (2008). http://dblp.uni-trier.de/db/conf/csse/csse2008-1.html#WuHL08

Wu G, Harrigan M, Cunningham P (2012) Classifying wikipedia articles using network motif counts and ratios. In: Proceedings of the eighth annual international symposium on wikis and open collaboration, WikiSym '12. ACM, New York, NY, USA, pp 12:1–12:10

Yeang CH, Huang LC, Liu WC (2012) Recurrent structural motifs reflect characteristics of distinct networks. In: Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012), ASONAM '12. IEEE Computer Society, Washington, DC, USA, pp 551–557. doi:10.1109/ASONAM.2012.94

Yu K, Lafferty J, Zhu S, Gong Y (2009) Large-scale collaborative prediction using a nonparametric random effects model. In: ICML. ACM, New York, pp 1185–1192