Community Evolution in A Scientific Collaboration Network

Minh Van Nguyen, Michael Kirley, and Rodolfo García-Flores

***Abstract*—A community in a network is a set of nodes with a larger density of intra-community links than inter-community links. Tracking communities in a network via a community life-cycle model can reveal patterns on how the network evolve. Previous models of community life-cycle provided a first step towards analyzing how communities change over time. We introduce an extended life-cycle model having the minimum community size as a parameter. Our model is capable of uncovering anomaly in community evolution and dynamics such as communities with stable or stagnant size. We apply our model to track, and uncover trends in, the evolution of communities of genetic programming researchers. The lifespan of a community measures how long it has lived. The distribution of lifespan in the network of genetic programming researchers is shown to be modeled as an exponential-law, a phenomenon yet to be explored in other empirical networks. We show that our parameter of minimum community size can significantly affect how communities grow over time. The parameter is fine-tuned to detect anomaly in community evolution.**

I. Introduction

Networks are ubiquitous in modern society. Over the past decade, complex systems found in social interactions, economics, science, and technology have been analyzed and modeled using the concept of network as a unified approach [1]. One challenge is to uncover trends in, and characteristics of, networks that change over time using special clusters of nodes called communities. In addressing this challenge, an objective is to discover principles underlying community evolution.

In a network, a community is a set of nodes having a larger density of intra-community links than inter-community links [2]. Tracking communities over time can uncover long-term trends in how communities evolve in the underlying networks. For example, in a blog network we might wish to detect which communities of blogs are relatively stable in size over a period of time [3]. In a mobile phone network, changes in community size over a timeframe can reveal calling patterns and customer churns [4], [5]. In other contexts such as scientific collaboration networks, communities of researchers that span many years suggest long-term research collaboration [5]. Such communities can be further investigated to uncover researchers in particular fields who are consistently productive over a period of time [3].

Previous work on scientific collaboration networks focused on evolution of the networks as a whole or individual node properties [6]–[10]. These studies presented local statistics such as clustering coefficients, number of collaborators per author, and average number of papers per author. Also considered were global statistics including size of the largest component, average path lengths, and degree distribution. A property shared by scientific collaboration networks is the small-world effect: high clustering coefficients and low average path lengths. Where communities were considered, as was the case for the network of genetic programming researchers [7], communities were extracted from the largest component of the network and little attempts were made to investigate community evolution.

An approach to track communities over time is to view community evolution in demographic terms. Previous work along this line analyzed community changes using a life-cycle model comprising events such as birth, death, expand, contract, merge, and split [4], [5], [11]. Unlike [4] and [5], Asur et al. [11] emphasized the life-cycle of nodes, an emphasis that is impractical in networks with millions of nodes and irrelevant when an overview of how communities evolve is required. Palla et al. [5] used the above events to quantify the evolution of a phone call network and a coauthorship network, whereas Greene et al. [4] used the events to investigate community evolution in a phone call network. Except for [5], little attention was paid to modeling an event as a function of time. Life-cycle models in the above work provided a first step towards analyzing community evolution. Issues that were not addressed include: Under what condition(s) would a change in community size be stable? To what extent does the minimum community size affect the evolution of communities?

In this paper, we introduce an extended life-cycle model to address issues such as the above, among others. Our model extends previous work by introducing a set of extra events and a parameter for the minimum community size. As a case study, we apply our model to the network of genetic programming (GP) researchers as documented in the GP bibliography.[1] This dataset is a comprehensive record of research publications in GP, hence an analysis of the dataset is expected to provide some general conclusions about the state of GP research. We show that trends in community events within the GP network can be modeled as functions such as exponential for birth, logarithmic for direct (an event in our model), and polynomial for expand. The lifespan of a community measures how long the community has lived since birth [5]. We show the distribution of lifespan within the GP

Minh Van Nguyen and Michael Kirley are with the Department of Computing and Information Systems, University of Melbourne, Melbourne, Victoria 3010, Australia (minguyen@student.unimelb.edu.au, mkirley@unimelb.edu.au).

Rodolfo García-Flores is with CSIRO Mathematics, Informatics and Statistics, Clayton, Victoria 3169, Australia (Rodolfo.Garcia-Flores@csiro.au).

[1] http://www.cs.bham.ac.uk/~wbl/biblio/

network to be modeled as an exponential-law, a phenomenon that to the best of our knowledge is yet to be explored in other real-world networks. Our parameter of minimum community size is shown to significantly affect, and can be used to detect anomaly in, community evolution.

The rest of the paper is organized as follows. Section II introduces our extended life-cycle model. In section III, we describe our experimental setup and parameter settings. Sections IV and V discuss results of our experiments in applying our extended life-cycle model to the network of genetic programming researchers. We conclude the paper in section VI.

II. Methods

Our extended life-cycle model is best understood in the context of time-stamped network datasets and the life-cycle model in [4] and [5]. We first describe how to preprocess a time-stamped network dataset, in preparation for tracking community evolution via our extended life-cycle model or the previous life-cycle model.

A. Preprocessing

A general procedure to preprocess a time-stamped network dataset is as follows. First, define the objects in the dataset to represent as nodes. Next, define the relationship to be used to connect two nodes via an edge. In some cases, the dataset might need to be cleansed to remove noise or disambiguate objects. A network snapshot at time t is historical as it contains all nodes and edges existing at or prior to t. Extract all communities in the snapshot via a community detection algorithm, e.g. an algorithm in section II-A4. By the end of the preprocessing stage, we have a set C_t of communities for the snapshot at t. All the C_t are fed in temporal order to an algorithm for tracking community evolution based on a life-cycle model, such as the algorithm in Figure 1. Below, we elaborate the procedure for the case of the GP dataset.

1) Nodes and edges: The objects of interest are author and, to a lesser extent, editor names. We define the collaborators of a publication as its authors and people who served as the publication's editors (if any). We derive two versions of the GP network: the network with only authors (the coauthor network), and the network including both authors and editors (the collaborator network). In [7], collaborators were defined as strictly coauthors or coeditors. Our definition of collaborator accounts for the roles that editors play in shaping the technical contents of papers in edited volumes. In summary, nodes represent people with names in the GP dataset. Two nodes are linked by an edge if the corresponding people have coauthored, coedited, or collaborated on a publication.

2) Data cleansing: In automated extraction of author and editor names from a bibliographic dataset, bias in the choice of name representation should be minimized. One approach to control bias is to construct two different networks for the same bibliography, such that the networks provide rough lower and upper bounds for various statistics on the original dataset [12]. Following [7], we use the names as given in the GP dataset. This is supplemented with minimal name cleansing. Each occurrence of a generational suffix is standardized to a lowercase equivalent without periods, e.g. "jr" or "sr". Where relevant, we insert missing periods after initials and remove white spaces between hyphens.

3) Snapshots: Each entry in the GP dataset is time-stamped by the publication year. We construct time-ordered snapshots based on author and editor names and the years of publication. The network snapshot for a year consists of all papers and volumes published before or during that year. The GP dataset has publications since 1950, thus we have yearly snapshots from 1950 up to and including 2011. Each snapshot is represented as an undirected graph without self-loops nor multiple edges. Nodes and edges are as defined above. We retain isolated nodes, since a node might be isolated in one snapshot but is connected with another node in a subsequent snapshot.

4) Communities: All communities in each snapshot are extracted via the algorithms in [13] and [14], each of which produces non-overlapping communities. The algorithm in [13] uses modularity optimization and hence suffers from the problem of resolution limit [15]. That is, it is possible for an algorithm using modularity optimization to combine multiple smaller communities into one large community when doing so optimizes its objective function. In contrast, the algorithm in [14] uses label propagation and is immune to such problem. These algorithms are chosen because their implementations are publicly available in the igraph[2] C library and they scale to the size of the entire GP network. Using different techniques to extract communities allows us to compare and contrast results based on various community detection algorithms.

B. Life-cycle model

The life-cycle of dynamic communities can be described via six events: birth, death, merge, split, expand, and contract [4], [5]. Before defining these events, we define dynamic communities and show how to capture their evolution as timelines.

Let Γ be a network whose snapshots are given as step graphs $G_0, \ldots, G_n$. Each snapshot G_t captures the state of Γ at time t. We want to identify a set of k' dynamic communities $D = \{D_0, \ldots, D_{k'-1}\}$, where each D_j is present in Γ at one or across multiple snapshots. In each G_t is a set of k_t step communities $C_t = \{C_{t0}, \ldots, C_{t(k_t-1)}\}$, where each C_{ti} is a snapshot at time t of some D_j. The evolution of D_j up to and including time t is represented as a timeline comprising of step communities $C_{t'i}, C_{(t'+1)i}, \ldots, C_{ti}$, where $0 \leq t' \leq t$. The most recent observation in the timeline of D_j is called the front of D_j, denoted F_j. The lifespan ℓ of D_j is the number of snapshots in which it exists. If $G_{t'}$ and G_t are the first and last snapshots in which D_j exists, respectively, then $\ell = t - t' + 1$.

1) Birth: Birth is the emergence of a new D_j distinct from any extant $D_i \in D$. At some time t, a step community C_{tl} emerges, but does not match the front of any $D_i \in D$. We create a new dynamic community D_j, let C_{tl} be its first step community, and add D_j to D. The number of birth in t counts how many dynamic communities first emerge in t.

[2]https://launchpad.net/igraph

2) Death: Death is the dissolution of $D_j \in D$ after a fixed number of consecutive time steps in which the front F_j does not match any step communities. Fix a death threshold $\delta > 0$ specifying the number of time steps prior to terminating D_j. If during at least δ consecutive time steps F_j does not match any step communities, then we remove D_j from D. The number of death in t counts how many dynamic communities have their very last observation in t.

3) Merge: A merge is the join of multiple dynamic communities into one. Let $D_{a_0}, \ldots, D_{a_n}$ be distinct dynamic communities. If at time t the fronts F_{a_i} match the same step community C_{tk}, then the D_{a_i} are said to have merged into one dynamic community. From t onwards, all D_{a_i} share the same timeline. The number of merge in t counts how many dynamic communities in t into which multiple dynamic communities from $t-1$ merge.

4) Split: A split is a branching of some D_i into multiple dynamic communities. That is, at time t, F_i is matched to multiple $C_{ta_0}, \ldots, C_{ta_k}$. We create new dynamic communities $D_{a_0}, \ldots, D_{a_k}$ sharing the same timeline with D_i up to $t-1$. The D_{a_j} have their own timelines starting from t onwards. The number of split in t counts how many dynamic communities in t into which dynamic communities from $t-1$ split.

5) Expand: A D_i is said to expand if its front at t has significantly more nodes than its front at $t-1$. Fix a growth threshold $0 < \gamma \leq 1$. Then D_i expands if its front at t has a minimum proportion of γ more nodes than the front at $t-1$. The number of expansion in t is the number of dynamic communities that expand in t from its previous size at $t-1$.

6) Contract: Contraction is a shrinking in size of some D_i. Let $F_{(t-1)i}$ and F_{ti} be fronts of D_i at times $t-1$ and t, respectively. Fix a contraction threshold $0 < \kappa \leq 1$. Then D_i has contracted if F_{ti} has a minimum proportion of κ less nodes than $F_{(t-1)i}$. The number of contraction in t is the number of D_i that contract from its previous size in $t-1$.

Based on the above events, Greene et al. [4] proposed an algorithm to track dynamic communities across time steps (see Figure 1). The algorithm takes as input $n+1$ time-ordered snapshots $G_0, \ldots, G_n$, a community detection algorithm K (e.g. see section II-A4), a similarity function S (e.g. the Jaccard coefficient), and a match threshold θ.

We have implemented the algorithm in Figure 1 using Python. The time complexity of the algorithm is dominated by the time complexities of K and S. The time complexity can be improved by extracting beforehand all communities from each snapshot graph. When executing each of lines 1 and 3, use the pre-extracted communities. This is the strategy we adopt. In our implementation, we first extract all communities using the igraph C implementation of the algorithms in [13] and [14] and then track dynamic communities using our Python implementation. We use the Jaccard coefficient to match communities. Two step communities C_{ti} and $C_{(t+1)j}$ in consecutive snapshots G_t and G_{t+1}, respectively, are said to match each other if their Jaccard coefficient satisfies

$$J(C_{ti},\, C_{(t+1)j}) = \frac{|C_{ti} \cap C_{(t+1)j}|}{|C_{ti} \cup C_{(t+1)j}|} > \theta.$$

Input: $G_0, \ldots, G_n$; K; S; θ
Output: Records of community matches.

```
1:  C_0 ← K(G_0);  D ← {D_i | D_i = C_0i ∀C_0i ∈ C_0}
2:  for t ← 1, ..., n do
3:      C_t ← K(G_t)
4:      for all C_ta ∈ C_t do
5:          match all D_i for which S(C_ta, F_i) > θ
6:          if no match exists then
7:              create new D_j and add to D
8:              let C_ta be the first observation of D_j
9:          else
10:             add C_ta to each matching dynamic community
11:         end if
12:         update front of each D_i to latest matched C_ta
13:         if some D_i matches multiple step communities then
14:             create a split dynamic community
15:         end if
16:     end for
17: end for
```

Fig. 1. Algorithm to track communities over time, as proposed in [4]. In this framework, K and S should be substituted for a specific community detection algorithm and a similarity function, respectively.

C. Extended life-cycle model

The life-cycle model above characterizes dynamic communities in a coarse-grained manner. Here, we introduce an extended life-cycle model that addresses finer-grained issues such as when the count of missing observations is within the death threshold, when the change in community size is neither a contract nor an expand, and so on. Our model incorporates the life-cycle model and enhances the latter with extra events: direct, missing, resume, stable, and stagnant. As noted in section I, in a network a community is generally agreed to be a set of nodes having a larger density of intra-community links than inter-community links. Beyond this description, we are not aware of a generally agreed upon formal definition that specifies exactly the minimum number of nodes a community must have. We address this issue by introducing the parameter $s_{\min}$, the minimum community size.

1) Direct: Let $C_{(t-1)a}$ and C_{tb} be communities in the timeline of D_j. If $C_{(t-1)a}$ matches only C_{tb} and vice versa, then C_{tb} is a direct continuation of $C_{(t-1)a}$. The count of direct continuation in t is the number of dynamic communities in t whose observations $C_{(t-1)a}$ and C_{tb} only match each other.

2) Missing: A D_j has missing observation in t if F_j does not match any step communities in t. The number of missing observation in t is the number of dynamic communities with missing observation during that time step.

3) Resume: A D_j has a resume observation in t if it has a missing observation at $t-1$, but F_j matches a step community at t. The number of resumption in t counts the number of dynamic communities in t with a resume observation.

4) Stable: Let $C_{(t-1)a}$ and C_{tb} be communities in the timeline of D_j, where the change in community size from $t-1$ to t is given by the proportion c. If $0 < c < \gamma$ or

TABLE I
STATISTICS FOR VARIOUS SCIENTIFIC COLLABORATION NETWORKS.

	GP1	GP2	Biology	Physics	Math
# authors	10,342	8,655	1,520,251	52,909	253,339
# papers	8,687	8,578	2,163,923	98,502	—
Papers/author	5.12	2.35	6.4	5.1	6.9
Authors/paper	5.90	2.43	3.75	2.53	1.45
⟨collaborators⟩	16.9	4.7	18.1	9.7	3.9
Largest comp.	73%	27%	92%	85%	82%
Mean distance	3.6	6.6	4.6	5.9	7.6
Diameter	13	17	24	20	27
Clust. coeff.	0.75	0.62	0.066	0.43	0.15
Assortativity	−0.14	0.96	0.13	0.36	0.12

$0 < |c| < \kappa$, then D_j has a stable community size in t. The number of stable dynamic communities in t is the number of dynamic communities with stable size in that time step.

5) Stagnant: Let $C_{(t-1)a}$ and C_{tb} be communities in the timeline of D_j, where the change in community size from $t-1$ to t is given by the proportion c. If $c = 0$, then D_j has a stagnant size from $t-1$ to t. The number of stagnant D_j in t is the number of dynamic communities with zero change in size from $t-1$ to t.

III. EXPERIMENTS

We use revision 1.1846 of the GP bibliography, dated 2011/08/24. This dataset is unusual due to its comprehensive record of research publications in GP. By analyzing the dataset, we expect to draw some general conclusions about the current state of GP research. As of August 2011, the GP collaborator network (see section II-A1 for definition) consists of 10,342 nodes and 87,704 edges. Exactly 299 nodes are isolated, i.e. people who have neither coauthored nor coedited a publication. The GP coauthor network has 8,655 nodes, 20,480 edges, and 776 isolated nodes.

Prior to tracking dynamic communities across snapshots, we set $s_{\min}$ to a positive integer and consider only those communities with size at least $s_{\min}$. In particular, we track all dynamic communities for each parameter setting $s_{\min} = 1, \ldots, 10$. We hypothesize that the parameter $s_{\min}$ does not affect any long-term trends in events of the life-cycle model and our extended life-cycle model. Following [4], we set the match threshold at $\theta = 0.3$, the death threshold at $\delta = 3$, the expand threshold at $\gamma = 0.1$, and the contract threshold at $\kappa = 0.1$. Whether we use our definition of collaborator (see section II-A1) or the definition in [7], we hypothesize that including editors can misrepresent the state of the GP network. The misrepresentation is expected to be manifested through the inflation of network statistics such as the size of the largest component, average path length, and the clustering coefficient. As the algorithm in [13] suffers from the problem of resolution limit (see section II-A4), we hypothesize that for some snapshots of the GP network the algorithm in [14] would output more communities than the algorithm in [13]. The dataset we use and source code of programs to process it are available at https://bitbucket.org/mvngu/cec2012-suppmat/.

IV. RESULTS: SUMMARY STATISTICS

This section and the next present results of applying our extended life-cycle model to the GP network. Here, we present an overview of the GP network and show that including editors can misrepresent the state of the network. In section V, we analyze the evolution of the GP network using our extended life-cycle model.

A. Summary statistics

Table I presents summary statistics for the GP and various scientific collaboration networks. GP1 and GP2 are the GP collaborator and coauthor networks, respectively. In the GP1 column, "authors" refers to collaborators as defined in section II-A1. In the GP1 and GP2 columns, "papers" refers to papers, books, and edited volumes. Biology is the Medline bibliography on biological research, Physics is the Physics E-print Archive, and Math is the network of mathematicians compiled from the *Mathematical Reviews* database. Values in the columns Biology, Physics, and Math are taken from [8]. A note of caution is in order concerning how the Biology, Physics, and Math datasets were preprocessed. Data for the Biology and Physics columns were computed from networks where each node maps to an author identified by their surname and all initials [8], [12]. Data for the Math column were derived from a network where each node represents an author name as it appears in *Mathematical Reviews*, except a few anomalous names (e.g. "et al.") were excluded [8], [16].

As is clear from Table I, including editors can inflate statistics on the whole GP network. Luthi et al. [7] noted a similar inflation, although their definition of collaborator is different from ours. The absolute counts of authors and papers are inflated, as are the average number of papers per author, the average number of authors per paper, and the average number of collaborators (denoted ⟨collaborators⟩). In GP1, the largest component includes about 73% of all nodes, resulting in reduced values for the average path length (mean distance) and length of the longest path (diameter). In terms of the average number of authors per paper and the average number of collaborators, the publication pattern for GP1 resembles that in biology than in either physics or mathematics. In contrast, the publication pattern for GP2 resembles that found in physics and mathematics. Both GP1 and GP2 exhibit the signature of small-world networks: relatively low mean distance as compared to the respective diameters, and high clustering coefficients (≥ 0.62). Thus, editors can misrepresent the state of the GP network by inflating various summary statistics.

Another way in which editors can misrepresent the state of the GP network is by deflating the mixing ratio (degree assortativity). GP1 is disassortative with mixing ratio $r \approx -0.14$ (see Table I). As collaborators of authors, editors (especially those with high numbers of collaborators) tend to connect with authors having relatively low numbers of coauthors, resulting in decreased path lengths in GP1. Note that this tendency is rather weak. With editors excluded (GP2), the GP coauthor network is highly degree assortative ($r \approx 0.96$). In this case, authors with high numbers of coauthors strongly tend

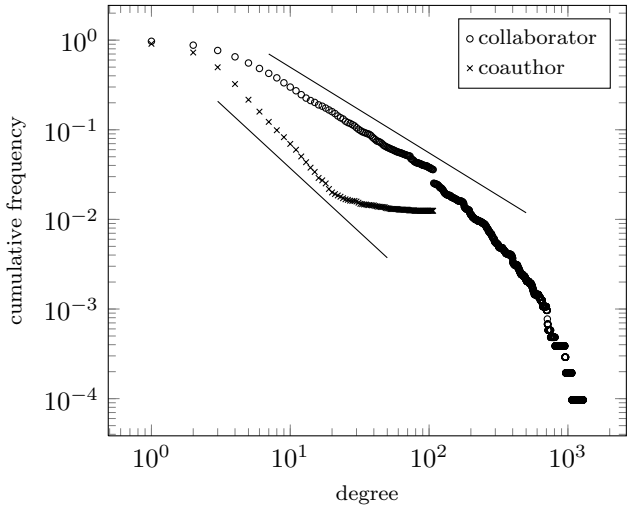


Fig. 2. Cumulative distributions of number of coauthors and collaborators. See section II-A1 for definition of collaborator.

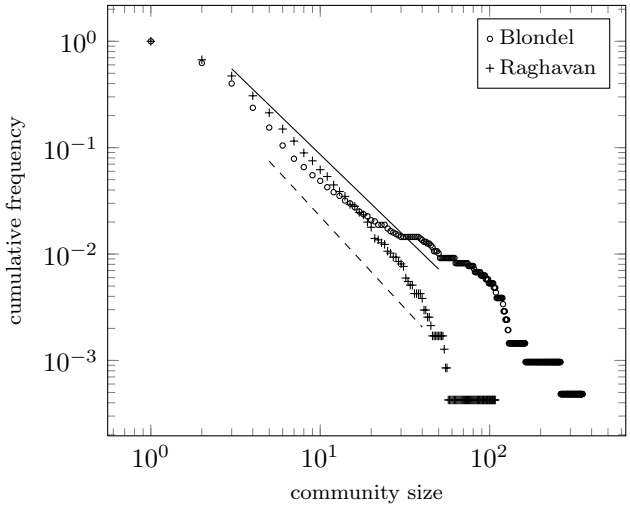


Fig. 3. Cumulative distributions of community size for the GP coauthor network. The distributions are derived from the 2011 snapshot. "Blondel" refers to communities extracted via the algorithm of Blondel et al. [13]. "Raghavan" refers to communities extracted via the algorithm of Raghavan et al. [14].

to collaborate with others having high numbers of coauthors. In contrast, authors with low numbers of coauthors strongly tend to collaborate with those of similarly low numbers of coauthors. Social networks found in scientific and professional collaborations are reported to be assortative, whereas technological and biological networks are disassortative [17]. Thus, GP1 is atypical of a scientific collaboration network. It resembles more of a technological or biological network, rather than a scientific or professional collaboration network.

B. Degree distribution

Both of the GP coauthor and collaborator networks cannot be modeled using pure power-laws of the form $p(x) \sim x^{-\alpha}$ for their degree distributions. Figure 2 shows the cumulative distribution of the number of coauthors (resp. collaborators) corresponding to the coauthor (resp. collaborator) network. The distributions are fitted with power-laws having exponents $\alpha \approx 2.42$ (coauthor) and $\alpha \approx 1.95$ (collaborator), both computed using the maximum likelihood estimator (MLE) technique in [18]. The power-law regimes are estimated to take effect from $x_{\min} = 7$ (collaborator) and $x_{\min} = 3$ (coauthor). Following the procedure of Clauset et al. [18], we use the Kolmogorov-Smirnov statistic to quantify the goodness-of-fit of the GP coauthor/collaborator networks to power-laws. The resulting p-values are < 0.05. Clauset et al. proposed the conservative rule that if the p-value is ≤ 0.1, then the power-law hypothesis is ruled out for the data in question. Using this rule, we conclusively reject that the GP coauthor and collaborator networks follow power-law degree distributions. Such phenomenon has been reported for related empirical networks. Luthi et al. [7], [10] reached the same conclusion for the GP dataset, but based on visual evidence in the plots of degree distributions. Above, we provided statistical evidence to support such conclusion. Other scientific coauthor networks are known to exhibit an exponential degree distribution or a power-law with an exponential cut-off [12], [16].

V. Results: community evolution

We present detailed analyses on community evolution in the GP network. As shown in section IV, including editors can misrepresent the state of the GP network, hence we focus exclusively on the GP coauthor network. Only snapshots since 1995 are considered, because it is from 1995 onward that the coauthor network has communities with $s_{\min} = 10$.

A. Distribution of community size

Here, we model the distribution of community size in the GP coauthor network as a power-law. Figure 3 shows the cumulative distributions of community size. Communities are extracted from the 2011 snapshot via the techniques in [13] (denoted "Blondel" in the figure) and [14] (denoted "Raghavan"). At first, the Raghavan distribution is higher than the Blondel distribution, but from size 15 onward the Raghavan distribution is below the Blondel distribution. This cross-over can be explained by the problem of resolution limit inherent in the algorithm of [13] (see section II-A4). As the latter algorithm uses modularity optimization, it tends to aggregate smaller communities into a large community, hence outputs a smaller number of communities than the algorithm in [14]. We note this for all snapshots since 1999, confirming our hypothesis from section III.

Using procedures in [18], the Blondel distribution has a power-law exponent $\alpha \approx 2.54$ (solid line in Figure 3) and p-value ≈ 0.123 for goodness-of-fit. The power-law regime is estimated via MLE to take effect from $x_{\min} = 3$ and we observe the power-law decay for community sizes up to $x \approx 50$. The Raghavan distribution has a slightly higher exponent ($\alpha \approx 2.72$, dashed line) and $p \approx 0.995$. We estimate the power-law regime to start from $x_{\min} = 5$ and observe the decay for all community sizes up to $x \approx 40$. Using the rule of Clauset et al. [18] for accepting/rejecting the

power-law hypothesis, since $p > 0.1$ for both of the Blondel and Raghavan distributions, we accept it is possible for the GP coauthor network to have community size distributed as a power-law. Furthermore, each snapshot of the coauthor network from 1995 to 2010 has power-law scaling in its community size distribution with $p > 0.1$. Thus, we have statistical evidence that community size in the GP coauthor network can be modeled as a power-law.

The above scaling behavior of community size distribution has been observed in other real-world networks. Various subsets of the arXiv dataset have been shown to have power-law exponents $\alpha \approx 0.54, 0.97, 1.07, 1.6$ in their community size distributions [19], [20]. Other empirical networks exhibiting power-law scaling in their community size distributions include: protein interaction and word association networks [20]; email interaction and jazz musician networks [19]; and an Amazon copurchasing network [21]. Such a phenomenon suggests that, analogous with networks having scale-free degree distributions, many types of real-world networks are heterogeneous in the distributions of their community sizes. To the best of our knowledge, an explanation for this phenomenon is yet to be proposed.

B. Trends in events

We model events in the life-cycle of dynamic communities as exponential, logarithmic, and polynomial functions. Events in our extended life-cycle model are shown to provide finer-grained analysis of how the GP network evolves than events in the life-cycle model. Due to limitation of space, we only discuss some trends apparent from grayscale maps (see Figure 4) of time series data on normalized counts of events.

Figure 4 shows grayscale maps for four events in the life-cycle of dynamic communities. All communities are extracted via the algorithm in [13]. We omit grayscale maps based on communities extracted via the algorithm in [14], as these are qualitatively similar to the maps derived from communities extracted via the algorithm in [13]. Also omitted are maps for other events described in section II. The maps can be read as follows. For each setting of the parameters $s_{\min}$ and year, we count the number of occurrences of an event and normalize that number by the count of dynamic communities for the given parameter setting. The resulting normalized count is shown in the grayscale map for the event under consideration. For a given year and a particular setting of $s_{\min}$, as the normalized count of an event approaches 1 (resp. 0), the higher (resp. lower) is the proportion of dynamic communities experiencing the event in question. A normalized count of, say, birth close to 0 indicates that a relatively small proportion of new dynamic communities were born during a given year and for a particular setting of $s_{\min}$ (see Figure 4(a)). A normalized count of, say, stagnant close to 1 indicates that during a given year and for a particular setting of $s_{\min}$, the vast majority of dynamic communities did not change in size from their respective sizes in the previous year (see Figure 4(d)).

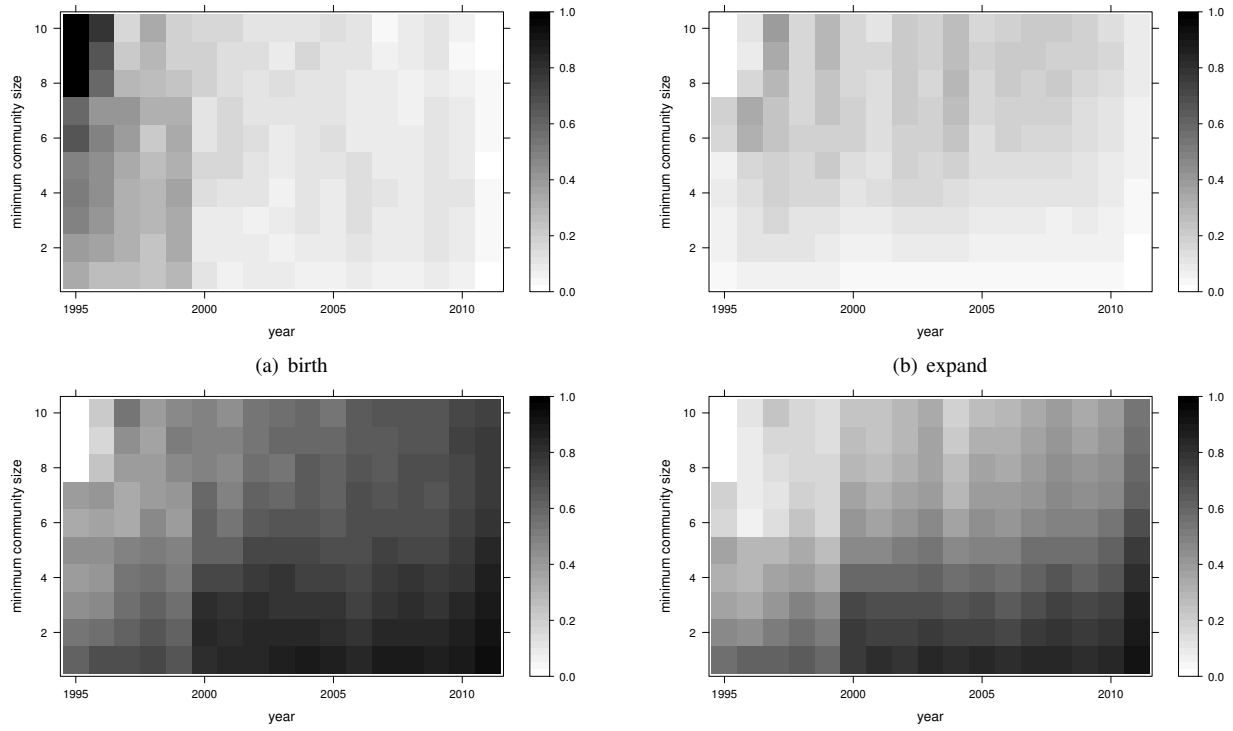
From Figure 4(a) note that over time the normalized count of birth, denoted b, decreases. The time series plots of b suggest an exponential decay. To confirm this, for each $s_{\min} = 1, \ldots, 10$ we use nonlinear least squares regression to fit the model $b \sim \exp(\alpha + \beta x)$ to the time series of b and obtain the bounds $0.02 < RSS < 0.17$ on the residual sums of squares. Thus, it is plausible that b decays as an exponential function.

Now consider how the normalized count of expand, denoted e, changes over time (see Figure 4(b)). Unlike b, e appears to follow two decay regimes depending on the parameter $s_{\min}$. For each $s_{\min} = 1, 2, 3, 4, 6, 7$, we use linear regression to fit the time series of e to the model $e \sim \alpha x + \beta$, with the null hypothesis $\alpha = 0$. The resulting p-values are < 0.05, so it is plausible that e decays linearly. However, the same procedure and null hypothesis result in p-values > 0.05 for $s_{\min} = 5, 8, 9, 10$. As an alternative, with the latter settings for $s_{\min}$, we use nonlinear least squares regression to fit the time series of e to the model $e \sim \sum_{i=0}^{5} \alpha_i x^i$. The residual sums of squares are bounded by $0.008 < RSS < 0.051$. Therefore, e appears to follow two decay patterns depending on values of $s_{\min}$, thus invalidating our hypothesis from section III.

The normalized counts d and s of direct continuation and stagnant size, respectively, both appear to increase over time (see Figures 4(c) and 4(d)). The time series plots of d suggest two growth trends: one is logarithmic and the other linear. For any $s_{\min} = 1, \ldots, 10$, we use nonlinear least squares regression to fit the time series of d to the model $d \sim \alpha + \beta \log(x)$, using the natural logarithm. The resulting bounds on the residual sums of squares are $0.02 < RSS < 0.067$. Consider again each of the above settings for $s_{\min}$ and suppose that $d \sim \alpha x + \beta$, where $\alpha \geq 0$. With the null hypothesis $\alpha = 0$, linear regression on the time series of d produces p-values < 0.05. With the same null hypothesis, linear regression on the time series of s also results in p-values < 0.05. Unlike the decay patterns of e, the minimum community size does not seem to affect the growth patterns of d and s.

The trends in the proportion of split suggest some interesting dynamics within the wider community of GP researchers. Irrespective of $s_{\min}$, it is only since 2001 that we have splits in dynamic communities, with normalized count < 0.036 for each year from 2001 onward. This is possibly a sign of healthy inter-disciplinary collaboration among GP researchers during the decade 2001–2011. Another hypothesis is that in the last decade, various research projects came to an end and the participants scattered themselves amongst other projects or started new collaborative research projects. Such projects may be within standard topics of genetic programming or more likely be centered around hot topics. It would be interesting to test such conjectures based on the GP bibliography, in tandem with anthropological or sociological analyses.

We now synthesize the trends apparent in Figure 4. Taking into account the exponential and linear decays of b and e, respectively, these suggest the presence of events other than those in the life-cycle model. Figure 4 suggests that the relatively low and decreasing proportions of events in the life-cycle model are to a large extent offset by the increasing proportions of events in our extended life-cycle model. The general pattern is that the relatively low and in some cases de-



(a) birth (b) expand

(c) direct (d) stagnant

Fig. 4. GP coauthor network: normalized counts of events in the life-cycle of dynamic communities for given settings of the parameters $s_{\min}$ and year. The top two grayscale maps are for events in the life-cycle model. The bottom two maps are for events in our extended life-cycle model. Each count is normalized by the number of dynamic communities for a given parameter setting. The counts are based on communities extracted via the algorithm of Blondel et al. [13].

creasing proportions of birth, death, expand, contract, merge, and split are offset by the increasing proportions of direct continuation and zero change in community size. Dynamic communities of GP researchers appear to experience relatively few split and merge, but a high number of direct continuation. Over time, a relatively large proportion of dynamic communities appear to be likely to have no new members from one year to the next. We further address this point in section V-D, where stagnant size forms part of a signature of anomaly.

C. Distribution of lifespan

We model the distribution of lifespan of dynamic communities as an exponential-law. The longest possible lifespan for any dynamic community in the GP network is 62, corresponding to the 62 network snapshots from 1950 up to 2011. Figure 5 shows the distributions of lifespan for the GP coauthor network, for each minimum community sizes $s_{\min} = 1, 4, 7, 10$. We only show distributions based on communities extracted via the algorithm in [13], but we obtain qualitatively similar distributions when communities are extracted via the technique in [14].

From Figure 5, note that the distribution of lifespan ℓ can be fitted with an exponential-law $p(\ell) \sim \exp(-\lambda\ell)$. For all combinations of community detection algorithm [13] or [14], and minimum community size ($s_{\min} = 1, 4, 7, 10$), we have the parameter bounds $0.10 < \lambda < 0.19$. In Figure 5, with $s_{\min} = 1$ and the algorithm in [13], we fit the distribution of ℓ to an exponential-law with $\lambda \approx 0.101$. A significant caveat in the cases $s_{\min} = 1, 4, 7, 10$ is that, when fitting the distribution of ℓ with a power-law, the power-law regime is estimated via MLE to take effect from $x_{\min} \geq 13$ for the algorithms in [13], [14]. In other words, using the algorithm in [13] we would ignore between 60% to 76% of all dynamic communities in order to fit a power-law to the remaining data. The corresponding percentage range when using the algorithm in [14] is 63% to 87%. The distribution of ℓ shows good fit to a power-law, but usually for $< 40\%$ of all dynamic communities. It seems plausible that an exponential distribution might provide a better fit than a power-law. To the best of our knowledge, the distribution of lifespan of dynamic communities is yet to be investigated in real-world networks other than the network of genetic programming researchers.

D. Detecting anomaly

Relatively small values of $s_{\min}$ can be used to detect anomalous dynamic communities. A signature of anomaly can be seen in dynamic communities with relatively long lifespans (e.g. over half the number of snapshots), but are consistently stagnant in size. This might be a sign of people leaving the wider community of GP researchers, or in the worse case scenario the death of a researcher. For example, we consider all communities extracted via the algorithm in [13] and set $s_{\min} = 1$. The dynamic communities consisting solely of George R. Price, R. M. Friedberg, and Alan Turing span

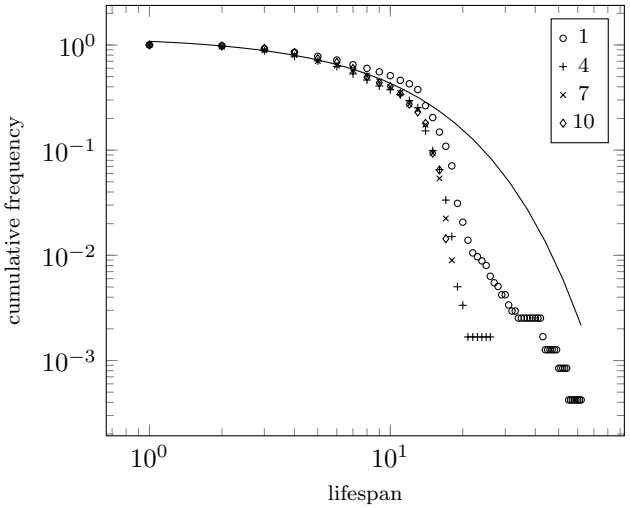


Fig. 5. Cumulative distribution of lifespan for dynamic communities in the GP coauthor network. Communities are extracted via the algorithm of Blondel et al. [13]. The distribution with $s_{\min} = 1$ is fitted with the exponential distribution $p(\ell) \sim \exp(-\lambda\ell)$, where $\lambda \approx 0.101$.

1970–2011, 1958–2011, and 1950–2011, respectively. All of these researchers, except for R. M. Friedberg, are known to be dead. Our parameter $s_{\min}$ can be used to separate historical dynamic communities from those with active researchers.

The example above illustrates shortcomings in how network snapshot and community are defined. Our definition of network snapshot views a snapshot as historical, so nodes representing dead people would remain in perpetuity. This issue was discussed in [10], where the notions of effective network and sliding time window were used to approximate the current state of collaboration within a given timeframe. Also, there lacks a generally agreed upon formal definition of community that specifies exactly the minimum community size. Together, the shortcomings above raise the following issue: To what extent do bibliographic databases capture the current state of research in particular disciplines?

VI. Conclusion

We have introduced an extended life-cycle model of communities that contains the parameter $s_{\min}$, the minimum community size. Using the network of genetic programming (GP) researchers as a case study, we have shown that our model allows for a finer-grained analysis than previous models on community evolution. The effect of $s_{\min}$ on community evolution depends on which community event we focus on. The parameter $s_{\min}$ did not seem to affect the trends in community events such as birth, direct, and stagnant. However, different values of $s_{\min}$ resulted in different trends for expansion. We addressed the issue of stability of community size by defining the stable event. The proportion of change $c \neq 0$ in community size is said to be stable if c lies strictly between the thresholds for contraction and expansion. Furthermore, we have shown that $s_{\min}$ can be fine-tuned to detect anomaly in community evolution. From the GP network, we uncovered that the lifespan of communities is distributed as an exponential-law. An avenue for future research includes using other real-world datasets to test our model and our exponential-law of lifespan.

Acknowledgment

This work used computing resources supported by US National Science Foundation Grant No. DMS-0821725. Data analysis was performed using igraph, plfit, R, and Sage.[3] We thank Andrey Kan and Jens Pfau for their comments.

References

[1] L. da Fontoura Costa, O. N. Oliveira Jr., G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, and L. E. C. Rocha, "Analyzing and modeling real-world phenomena with complex networks: a survey of applications," *Adv. Phys.*, vol. 60, pp. 329–412, 2011.

[2] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, pp. 75–174, 2010.

[3] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng, "FacetNet: A framework for analyzing communities and their evolutions in dynamic networks," in *WWW*, J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, Eds., 2008, pp. 685–694.

[4] D. Greene, D. Doyle, and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in *ASONAM*, N. Memon and R. Alhajj, Eds., 2010, pp. 176–183.

[5] G. Palla, A.-L. Barabási, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, pp. 664–667, 2007.

[6] A. L. Barabási, H. Jeong, Z. Náda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Phys. A*, vol. 311, pp. 590–614, 2002.

[7] L. Luthi, M. Tomassini, M. Giacobini, and W. B. Langdon, "The genetic programming collaboration network and its communities," in *GECCO*, H. Lipson, Ed., 2007, pp. 1643–1650.

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

[9] M. Perc, "Growth and structure of Slovenia's scientific collaboration network," *J. Informetrics*, vol. 4, pp. 475–482, 2010.

[10] M. Tomassini and L. Luthi, "Empirical analysis of the evolution of a scientific collaboration network," *Phys. A*, vol. 385, pp. 750–764, 2007.

[11] S. Asur, S. Parthasarathy, and D. Ucar, "An event-based framework for characterizing the evolutionary behavior of interaction graphs," *ACM Trans. Knowl. Discov. Data*, vol. 3, p. 16, 2009.

[12] M. E. J. Newman, "Scientific collaboration networks: I. Network construction and fundamental results," *Phys. Rev. E*, vol. 64, p. 016131, 2001.

[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. 2008, p. P10008, 2008.

[14] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Phys. Rev. E*, vol. 76, p. 036106, 2007.

[15] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *PNAS*, vol. 104, pp. 36–41, 2007.

[16] J. W. Grossman, "The evolution of the mathematical research collaboration graph," *Congressus Numerantium*, vol. 158, pp. 201–212, 2002.

[17] M. E. J. Newman, "Mixing patterns in networks," *Phys. Rev. E*, vol. 67, p. 026126, 2003.

[18] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, pp. 661–703, 2009.

[19] A. Arenas, L. Danon, A. Díaz-Guilera, P. M. Gleiser, and R. Guimerá, "Community analysis in social networks," *Eur. Phys. J. B*, vol. 38, pp. 373–380, 2004.

[20] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814–818, 2005.

[21] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, p. 066111, 2004.

[3] https://github.com/ntamas/plfit, http://www.r-project.org, sagemath.org