Community-Attribute Models for Bibliographic Reference Information via Dynamic Graph Evolution

A THESIS PRESENTED
BY
ROSS RHEINGANS-YOO
TO
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
ARTIUM BACCALAUREUS
WITH JOINT CONCENTRATION IN THE SUBJECTS OF
COMPUTER SCIENCE AND MATHEMATICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL 2016

Revised July 2016; original copy available
via Digital Access to Scholarship at Harvard.

Thesis advisor: Margo Seltzer Ross Rheingans-Yoo

Community-Attribute Models for Bibliographic Reference Information via Dynamic Graph Evolution

Abstract

We present CAMBRIDGE, the first technique for evolutionary analysis of dynamic graphs via a community-attribute graph model. Community-attribute models have been shown to be superior to models conventionally used for evolutionary analysis, particularly in modeling community structures in networks where communities exhibit dense overlaps. Thus, our use of a community-attribute model for analysis of a bibliographic network evolving in time allows us to observe not only the evolution of discrete clusters, but also the evolution of the 'core' of nodes that are strongly linked to multiple communities simultaneously.

In particular, our approach allows us to observe and quantify how the sibling communities resulting from community-splitting events share and compete for external intercommunity influence inherited from parent communities. We present evidence that indicates that in such splitting events, highly-connected nodes that were part of the parent networks 'strong intercommunity ties' become concentrated in the siblings' intersection, whereas highly-connected nodes that are part of 'weak intercommunity ties' are dispersed to the individual sibling communities.

We discuss the implications of our findings for the field of evolutionary graph analysis and address the evident promise of dynamic community-attribute models in providing fully generative models for dynamic networks.

Contents

List of Figures

to M. H. Yoo:

scholar, scientist, teacher, grandfather.

Acknowledgments

I AM UTTERLY HUMBLED by the generosity and kindness of the community of advisors and dear friends that has supported me thus far through my undergraduate studies at Harvard. I remain deeply indebted to Daniel Margo for sharing his academic passion for the theory of the structure of networks with an eager undergraduate of 2013. His advisor and mine, Margo Seltzer, likewise deserves far more thanks than I can here deliver; without her firm encouragement, gentle support, and cutting insight, this thesis would not have reached fruition. As her student, research assistant, teaching fellow, and advisee I am fortunate to have been able to learn so much.

My other esteemed reviewers and esrstwhile professors, Michael Mitzenmacher and Yaron Singer, likewise have my duly deserved thanks, both for their service as thesis readers and for their excellent classes over the years. And I am extremely grateful to the kind friends who have served as reviewers in a far less formal capacity, without whom the foregoing document would bear far more mortifying errors than it does now: Daniel Fu, Scott Kominers, and Evan Zimmerman. Chief among this list are my first professors and most ardent supporters, Penny Rheingans and Terry Yoo, in whose footsteps I can only hope to follow.

Special thanks are due to Scott Kominers and Christina Teodorescu for patient help refining my attempts at academic humor. And for their unfailing moral support in the past months and weeks, I thank Miriam Barnum, Keno Fischer,

David Holland, Alexander Lombardi, Christopher Merchantz, Lucian Wang, Cyndia Yu, and Ava Zhang.

Finally, I have no words at all to express the depth my gratitude to Sung Ja Yoo and the late Man Hyong Yoo, whose constant encouragement and unwavering support has followed me for twenty-two years and without which little—if any—of this would have come to be.

The first challenge for computing science is to discover how to maintain order in a finite, but very large, discrete universe that is intricately intertwined.
And a second, but not less important challenge is how to mould what you have achieved in solving the first problem, into a teachable discipline.

Edsgar W. Dijkstra

1

Introduction and Inquiry

How do populations form connections? How do those connection structures evolve as the populations grow?

The nascent field of network structure theory has endeavored to address these questions—with approaches that have shifted dramatically as new computational and mathematical technology has enabled algorithmic study of network patterns observed in the real world.

While some computer scientists have sought to develop expressive models for describing the structures observed in graphs, others continue to investigate the processes by which such structures themselves arise from the local decisions of individual agents. Our work extends this latter project, presenting a model for the ways in which 'communities' within networks appear, evolve, and divide into new sub-communities.

1.1 What we mean by "networks"

People form friendships and communities emerge…

Academic researchers work together and specialize into fields…

Autonomous systems form Internet relays and form a spider's-web of Internet backbone…

Pages on the Web link to one another, some become popular, others languish, and an entire industry arises to discover how they all relate to one another.

In the past fifty years, a convergence of scientific interest and engineering developments has enabled computer scientists to study observational data regarding the relationships *among* agents, systems, or texts in their own light. Interest in the structure of *networks*—abstractions of agent relationships as graphs of *vertices* connected by *edges*—dates at least as far back as Paul Milgram's 1967 *small-world experiment*, which found that, in 35 cases out of 160, a letter mailed to a random resident of Wichita, Kansas or Omaha, Nebraska could be forwarded through a chain of personal acquaintances to a stockbroker in Sharon, Massachusetts [51, 64].

Today, the study of networks continues fast apace, with analytic tools significantly more sophisticated than postcards and the kindness of strangers. A nuanced picture has emerged of features commonly observed in networks—as well as key axes of variation—in an incredible array of contexts: data collected on hyperlink networks in the World Wide Web [12, 37]; the spread and distribution of electronic files [23, 48] and email spam [18]; infrastructural networks for both electrical power [8] and Internet [25, 50]; collaboration and citation in academia [8, 43], the film industry [4, 8], and patent applications [43]; human interactions in online social networks [7, 45, 71]; biology and neurology [5, 66]; and a wealth of other environments [5, 55, 61].

Speaking broadly, we concern ourselves with networks that commonly exhibit a few characteristic features:

- **Degree distribution.** A few nodes (sometimes termed the *core*) are 'popular', having a large number of neighbors. Most, however, exist on a (relatively-)sparse *periphery* [3, 8].

- **Locality.** Nodes generally are members of small, closely-knit communities [45, 66].

- **Small and shrinking diameters.** Despite strong locality, randomly-selected pairs of nodes are surprisingly close to each other, on average. As networks grow in the number of nodes, the effective[1] diameter generally *shrinks* [17, 44].

- **Robustness.** The above three effects are robust to the removal of the nodes that most contribute to them. This suggests that they are not the result of a few outliers; rather, they are deeply embedded into the overall structure of the graph [11, 13, 16].

While these properties are sometimes referred to as 'graph laws', it is important to note that they are not so formal as that term suggests. Rather, the ways in which networks may vary in their adherence to each 'law' are often enlightening:

- Granovetter's 1973 "The Strength of Weak Ties" argued for the structural importance of single 'long-distance' links, in addition to neighborhood-based locality structures, and the importance of accounting for them in accurately modeling economic dynamics of networks [27, 28].

- Pennock, *et al.*'s 2002 "Winners Don't Take All: Characterizing the Competition for Links on the Web" challenged the understanding that the Web graph was dominated by its fat tail, suggesting that the global power-law distribution only arises as a mixture of lognormal distributions on indivdual subnetworks [53, 60].

[1] The *diameter* of a graph is the maximum distance between a pair of nodes; the *effective diameter* is a distance such that some majority of node-pairs are at most that distance apart.

- Boldi, Rosa, and Vigna, in 2012, suggested that the structural robustness of graphs varies significantly between two main types of graphs, indicating a 'web-like' *vs.* 'social-like' dichotomy [7, 15]. Social-like graphs (such as those observed in social, informational, and bibliographic networks), they suggest, are distinctly more robust in metrics like short-path-connectedness [11], and tend to exhibit multiple, overlapping layers of structure [14].

As we explore the structure of networks, we must thus take care to retain the proper perspective—informed by the general character of networks, yet aware that our intuition and heuristics may often be mistaken in particular cases that are more complicated than we first realize. Indeed, the twofold task—firstly, of formally characterizing the heuristically-known and secondly, of intuitively understanding the empirically-observed—remains the essential challenge in the study of networks.

1.2 Project overview

We take as the primary object of our study a bibliographic network that connects authors if one has cited the other in an academic paper in the ACM Digital Library. After cleaning and pruning poorly-connected authors from our dataset, we are left with 318,000 authors appearing in 293,000 publications published between 1951 and 2014, with 8,730,000 author-to-author links.

To decompose the latent structure of this network, we use the *community-attribute graph model* (AGM), a powerful new network model presented by Yang and Leskovec in 2014 [71], which labels nodes with membership in any number of 'communities', each of which exhibits a higher-than-average density of connections between member nodes

The primary innovation of the AGM over other community-membership network models (discussed later in §2.3.1) is in conceptualizing community memberships as additive *attributes*, rather than in a normalized *mixture*, so that additional community memberships do not detract from existing affiliations. By contrast, models that assign nodes mixtures of communities force nodes with joint

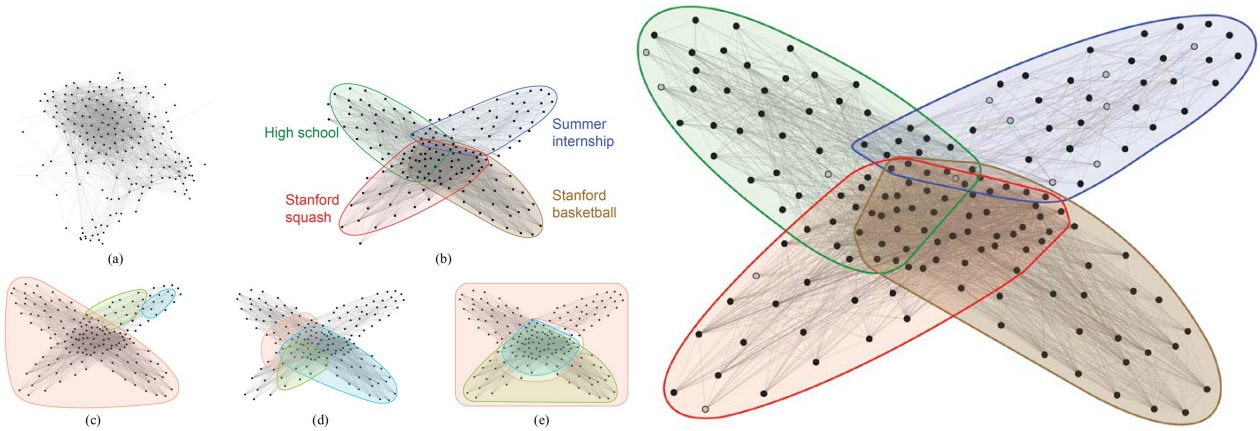

Figure 1.2.1: Figures from Yang and Leskovec [71]. **Left: (a)** a Facebook user's local friendship graph; **(b)** hand-labeled ground truth dividing it into four overlapping communities; **(c-e)** results of applying common clustering methods: **(c)** clique percolation **(d)** link clustering **(e)** mixed-memberhsip stochastic block. **Right:** the same graph with an AGM algorithmically fit by Yang and Leskovec's method.

membership to split their connections among multiple communities, remaining peripheral in each [70].

This innovation allows the AGM to better model networks where dense overlaps between networks are an important feature of the network structure. Consider the comparison between an AGM and various community mixture models on a graph taken from a Facebook user's local friendship graph in Figure 1.2.1, where only the AGM correctly interprets the dense core of the network as an overlap between multiple communities—sorting core nodes according to their signatures of external connections to the periphery structure—rather than grouping them into a dense super-community. We hypothesize, then, that the AGM will also be useful to describe the structure of an academic citation network, where we likewise wish to understand the fine structure of the network core in terms of author's affiliations with overlapping topic-based community clusters.

To develop a *dynamic* AGM, we take 'snapshots' of this network at regular intervals—considering only papers published before a given date—and fit an as-

signment of communities to each network with a statistical approach presented by Yang and Leskovec [68, 69]. With a combination of established techniques in evolutionary analysis [22] and techniques we develop specifically for analysis of AGM communities, we compare the community structure at each snapshot with the structure fit to the snapshots immediately before and after it, observing both the changes within a single community and flow of members between communities.

1.3 Contributions

We present the first techniques for fitting a dynamic AGM to a network evolving in time and the first dynamic network model that is able to properly represent dense overlaps between communities. Our techniques synthesize both methods for fitting AGMs to static graphs, adapted methods from other evolutionary analysis models, and specific innovations required by the community-attribute model.

Thus, we are able to present the first evolutionary analysis of *intercommunity* structure in networks that properly models the core–periphery and community-overlap structure observed in social and social-like networks. We present findings regarding the dynamics of 'inheritance' of external intercommunity relationships during events when one community splits into two or more over a series of snapshots.

Our work suggests a novel paradigm for generative graph models, namely, treating AGM-type communities themselves as node-objects interacting in a growing meta-network. Such a two-layer model would provide the first truly generative model for attribute-label graphs by simultaneously growing the network of communities and continually populating it with new nodes and internal connections. We conclude with a discussion of the obstacles remaining to the development of such a model and how they might be surmounted by further inquiry using our analytic techniques as well as others yet to be developed, building on our work.

What is true of one apple may not be true of another apple; thus more can be said about a single apple than about all the apples in the world.

Eliezer Yudkowsky

That's all very well in practice, but how does it work in theory?

aphorism, of disputed provenance

2

Motivation and Background

The work we present sits at the intersection of two long-established lines of inquiry: *What is the 'right way' to model network structure?*, and *How do networks evolve over time?* Though the two questions regard the same object of study, the first is a descriptive exercise that aims to interpolate between observed facts, while the latter is a predictive one, seeking to extrapolate network-structural 'laws of motion'.

This chapter situates our work in the arc of each question and explains the particular suitability of the *attribute-label* paradigm[1] for our purposes. In so doing, we provide a limited technical survey of relevant models, briefly presenting their relative strengths and weaknesses.

[1] Recall that *attributes* (or labels) can be layered additively, whereas normalized *mixtures* force nodes to trade off one affiliation for another.

2.1 A history of network models

Consider, as a source of intuition, the network connecting Harvard undergraduates if they have exchanged email messages in the past year. While any one *specific* graph might be able to provide information about a given population of students, we might instead seek to investigate the dynamics of email correspondence *in general* by viewing the variation between nodes as the emergent result of some general, generative process.

To emphasize this distinction, we borrow terminology from the Stanford Network Analysis Project (SNAP) [42], distinguishing *graphs*—the data of particular, fixed vertex-edge topologies—from *networks*, general models for connection dynamics arising from the interactions of node-local properties.

We might expect, for example, gross structural phenomena observed in one students' email network to be largely similar in most related networks, *e. g.*, other years' or schools' students/emails networks, other samples of communication between college students, or even observations of more general communication patterns among agents in other settings.

In the sections below, we will use this network to illustrate various historical models proposed for modeling networks. In so doing, we hope to provide intuitive motivation for the model that we select for use later in this work. Analogous examples include:

- Friendships registered on a social networking site.
- Cross-references in an online encyclopedia.
- Sets of products commonly purchased together through an online marketplace.
- Citations, collaborations, or other relatedness measures in document corpora, such as patents, legal opinions, or academic publications.

Remark 1. *In some settings, this list might be extended to include networks of hyperlinks on the World-Wide Web, connections between Autonomous Systems in the Internet, metabolic and protein-protein interactions, or ecological foodwebs. However, some authors have suggested that these networks are structurally distinct from the foregoing,*[2] *differing in certain important structural properties [7, 11, 13]. We defer the discussion of this distinction until §2.3.2 below.*

The sections that follow provide a brief historical survey of methods for network graph generation.[3]

An extremely simple model of this email network might assume that students send emails to random recipients at some constant rate. Thus, any given pair of students will be connected with some probability p, independent of how many other connections each has, and how many connections they share.

2.1.1 Gilbert, Erdős, and Rényi

Historically speaking, random graph models date back to the uniform models proposed by Gilbert [26], Erdős, and Rényi [24] for use in probabilistic proof methods. Probabilistic methods, broadly speaking, exploit statistical analysis of a distribution $\widetilde{\mathcal{G}}$ over a family of graphs $\mathcal{G}$ to demonstrate combinatorial facts:

- If the expected number of features b in a $\widetilde{\mathcal{G}}$-randomly-selected graph in $\mathcal{G}$ is less than 1, then *some* graph in $\mathcal{G}$ has no bs:

$$E\big(b\big(\widetilde{\mathcal{G}}\big)\big) < 1 \Longrightarrow \exists G \in \mathcal{G} : b\big(G\big) = 0. \tag{2.1}$$

- If the $\widetilde{\mathcal{G}}$-probability that a graph in $\mathcal{G}$ has the property A is greater than 0,

[3] A version of this discussion previously appeared in our survey of graph-generation algorithms [61].

then *some* graph in $\mathcal{G}$ has the property A:

$$\Pr\left(A\left(\widetilde{\mathcal{G}}\right)\right) > 0 \implies \exists G \in \mathcal{G} : A\left(G\right). \tag{2.2}$$

As statistical statements, these claims border on the tautological, but when applied to simple, analytically tractable distributions, they provide a bridge from statistical results to existential proofs. They are rendered most useful by the simple probability distributions to which they are usually applied:

- The *Erdős–Rényi exact-edge-count model* $\widetilde{\mathcal{G}}_{n,N}$ is equidistributed over all graphs on n vertices with N edges.
- The *Gilbert model* $\widetilde{\mathcal{G}}_{n,p}$ has n vertices, each pair of which is connected by an edge with independent probability p.

Remark 2. *Statistically speaking, the two are quite similar—note that the Gilbert model is distributed as a mixture of Erdős–Rényi exact-edge-count models with the edge-count distributed binomially as* $\mathrm{Bin}\left(\binom{n}{2}, p\right)$. *Since the binomial approaches zero relative deviation from its mean in the* $n \to \infty$ *limit, it is often useful to use the Gilbert model as an approximation, even of graphs of known edge-count. (Among its useful features is the fact that events involving disjoint regions of the graph are independent, greatly facilitating statistical feature-counting.)*

Indeed, since Erdős himself so often used the independent-edges formulation of the Gilbert model, the Gilbert model is often called "the Erdős–Rényi model", or ER. In agreement with the literature, we adopt this terminology, and refer to "exact-edge-count ER" explicitly when required.

The ER model, strictly speaking, is a model for *graphs*, though it is used remarkably often as a single-parameter model for randomly-generated networks when structural accuracy is not a significant desideratum. Again, it corresponds

to the case where connections between nodes are added independently and uniformly at random. Thus, the model has only two parameters: network size n and global density p.

2.1.2 The Structure of Networks

However, our experience with both email and college students might lead us to expect that our Harvard email network exhibit structural properties that this uniform model cannot represent well: interpersonal variation in email connectedness, densely-connected clusters of nodes, the existence of long-distance 'bridges' (either single nodes, or dense clusters) that prevent the diameter of the graph from growing, even as the number of nodes increases, *etc.*

We can quantify these phenomena, and note that they are exponentially unlikely even in ER models with a best-fit density p:

- **Some nodes are much more popular than others.** Node degrees in an ER network are distributed in a Binomial[4] distribution, but communication networks exhibit 'fat-tailed' behavior [3], in which highly-connected nodes are much more common than can be explained by a well-fit binomial distribution. (For further discussion of node-degree distributions, see Appendix B.)

- **Connected nodes share many neighbors.** **(1)** Any node's *neighborhood*—the set of nodes to which it is connected—has internal edge density significantly higher than p and **(2)** neighbor-pairs frequently exhibit *strong ties*, sharing a fraction of neighbors significantly higher than p [66].

- **Diameters are small, and *shrink* as networks grow.** The *graph diameter*—the largest distance between a connected pair of nodes—is smaller than the $O(\log n)$ predicted for ER graphs, and in fact, often remains stable or shrinks as graphs grow in the number of nodes [51].

To this set we often add a meta-property:

[4] A Binomial distribution, informally speaking, is a 'discretized Normal'.

- **The above three effects are largely robust.** None of these effects has a 'single point of failure', either a single outlier node or a small anomalous group. Even under the removal of the nodes that most contribute to node distribution, local-connectedness, or stable diameter, such phenomena tend to disappear gradually, rather than suddenly. This suggests that they are deeply embedded into the overall structure of the graph, rather than the result of a few outliers [11, 13].

  In our email analogy, robustness corresponds to believing that the structure we quantify in the foregoing ways generally reflects organic social dynamics rather than top-down administrative design.

If we are to understand the growth and structure of networks, we must be able to characterize, explain, and parametrize these phenomena (and their statistical subleties). It is this task that has given rise to the field of *network structure theory* as a distinct field of inquiry within computer science.

Remark 3. *Graph analysis often requires techniques distinct from 'tabular' data analysis (the conventional treatment of 'big data' as tuples of properties with static—if unknown—meanings). Traditional statistical methods for tabular data are often unable to properly treat 'rows' as column labels too, in settings where proper columnar attributes are of limited interest.*

Instead, it is often necessary in graph analysis to treat the relationships between *and* among *nodes as first-class objects—and since the space of possible 'relationships' is combinatorial in nature, simplification and abstraction are of utmost importance.*

2.1.3 Structural Approaches

As the first (often informal) network-structural "laws" were discovered, authors often sought to understand them through *ad hoc* models to simulate their structural

phenomena: Aiello, *et al.* modeled graphs with fat-tailed node degree distributions by considering an equidistribution over graphs with a suitable distribution of node degrees, *a la* the Erdős–Rényi exact-edge-count model [3]. Watts and Strogatz considered locality through a *regular ring lattice*—in which nodes in a circle are linked to their k nearest neighbors—augmented by uniformly-distributed 'long-distance' ER-like edges [66].

However, these models are no more plausible than ER as explanations of *why* large networks exhibit the structure that they do. (It is unlikely that variation in email connectivity arises only from a distribution of social popularity coefficients, or that it is dominated by the geographic layout of students.) For that task, we seek *generative* models, which allow observed phenomena to emerge as a result of local processes, rather than globally prescriptive structures.

Sequential-Attachment Models (1999-2005)

We might begin to enrich our email network with a notion of personal *popularity*. (A natural one, in which both engineers and members of the student computing society score quite highly.) Rather than passing judgment on students' personal qualities, however, we consider popularity *as reflected in the network graph* to determine a student's likelihood of acquiring a new connection. For example, we might imagine that any student forming a new contact is more likely to link to a student who has relatively many existing contacts, rather than one with relatively few.

Thus we arrive at the classic generative graph model, *BA*, originally presented by Barabási and Albert in 1999 as a technical implementation of the social maxim "the rich get richer" [8]. It was the first of the family of *sequential preferential attachment* models, wherein nodes are added iteratively to some small base graph, each choosing some m nodes to connect to by considering their respective degrees (as well as other considerations, potentially).

The original BA model uses strict *linear preference*—targets are chosen with probabilities proportional to their current degree. This induces a power-law distri-

bution:

$$p(\text{density} = d) \propto d^{\gamma} \tag{2.3}$$

with exponent $\gamma = 3$ [8], which can be adjusted to any desired $\gamma \in (2, 3)$ by introducing random re-wiring and the addition of extra edges between existing nodes[5] [4]. (Related models for node degree distributions are discussed in Appendix B.) Other authors, attempting to more closely fit empirical distributions of edge sources and targets, proposed affine-linear preference functions—some with positive [60], and others with negative [35], offset.

Still others, attempting to extend the model to include locality structures, proposed an alternative preference mechanism intended to induce community structure and power-law popularity simultaneously: Students are more likely to connect with the friends of their friends than with arbitrary strangers. In the family of *copying-based BA* models, new nodes select a *prototype*, then prefer to make connections to neighbors of their prototype, relative to other choices [37, 38].

This mechanism naturally induces a power law in node degree—since nodes with high degree are more likely to end up in the neighborhood[6] of a prototype—but, more importantly, makes it more likely for two connected nodes to share neighbors. More sophisticated copying-based models instead occasionally add selected nodes in the neighborhood of a prototype as prototypes themselves [44], providing a generative process that quite plausibly reflects how we expect social, bibliographic, and other informatic networks to grow.

We might even believe that these models, in some sense, shed light on what's 'really happening' in our email network at the level of node-local mechanics—we can model the precise extent to which the targets of newly-added links are likely to be nodes that are already well-connected, popular among the source node's existing friends, *etc.*

However, the family of increasingly-complicated BA variants began to face a

[5] Re-wired edges retain one endpoint and have the other re-assigned preferentially; extra edges have a uniformly-chosen source and a preferentially-chosen target.

[6] The *neighborhood* of a node is the set of nodes to which it is connected.

troubling practical difficulty. While the sequential-attachment model tells a compelling story about how global patterns emerge from local decisions, it is often mathmatically infeasible to marginalize the resulting distribution, to ask even simple questions such as "Well, what is the degree distribution, then?" While experimental evidence could shed some light on the functional dependencies, the resulting state of affairs was considered troubling to many who had hoped that models would give us tools amenable to mathematical—as well as empirical—analysis. Instead, the field was dominated by a procession of *ad hoc* BA variants whose fundamental mathematical structures could only be understood by peering through experimental keyholes.

Matrix Models (2004-2010)

In 2004, a research group from Carnegie Mellon proposed the *Kronecker-product model* (KrΠ) for networks which posited a fractal, nested community structure [19]. While it is often described as *recursive* in structure (indeed, the first versions were called *R-MAT*, for **R**ecursive **MAT**rix), this is merely an artifact of the adjacency-matrix representation; the actual structure is more properly conceptualized as a product of the initiator matrix in *k independent* indices.[7]

A *generalized* Kronecker-product (GKrΠ) model might describe a college email network as follows:

- Each student has a 'year'—students from the same year are more likely than average to be connected, and students within one year of each other slightly less so. Freshman–senior connections are the rarest. An equal number of students are in each year. (Years or year-pairs may have different densities, and the densities between pairs of years need not depend on the inter-year densities in any particular way.)

- Exactly 40% of students study 'science'; exactly 60% study 'humanities'; and within each year, these ratios are exact as well. (The number of students

[7] For an exposition of the matrix interpretation as well, see the chapter "Kronecker Graphs" in Jure Leskovec's CMU PhD dissertation [40].

in each year is, in fact, divisible by 5.) Each of science–science, science–humanities, and humanities–humanities has its own density of connections. The proportions between them hold when considering any year alone, or even any inter-year combination.

- Each student either lives on the 'north side' or the 'south side' of campus. For every student living on the north side, there are three students living on the south side in the same year and studying the same thing. North–north connections are the densest, followed by south–south, and finally by cross-campus connections—these densities, once again, remain proportional among any intra-/inter-year (and/or intra-/inter-subject) subsample.

All effects on connection density apply *multiplicatively*, and so to determine the connection probability between any pair of students, it suffices to know the year, subject, and location of each; then we simply product the year–year, subject–subject, and location–location interactions.

Furthermore, the independent and proportional distribution of types in each index allows us to arrange the node-to-node connectivity matrix (representing the probability that any pair is connected) into sixteen blocks, all identical up to a respective scalar factor, and each being further subdivided into twenty-five sub-blocks, again identical up to scalar factors (which are chosen from the three science/humanities density factors), and which can be further subdivided into sixteen uniform sub-sub-blocks to represent the north/south structure.[8]

In actuality, Leskovec, *et al.*'s Kronecker models (KrΠ) use repeated application of a single 'initiator matrix', rather than separate matrices for years, subjects, locations, *etc.* Formally:

[8]Since the effects of each feature are multiplicative and features have no particular order, we could just as easily divide the matrix by subject first, then location, then year.

- Let the initiator matrix A be some arbitrary $n \times n$ matrix:

$$A := \begin{pmatrix} a_{1,1} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \dots & a_{n,n} \end{pmatrix}. \tag{2.4}$$

  Each index of the initiator represents a node 'type'; each $a_{\alpha,\beta}$ represents the relative density[9] of connections from type α to type β.[10]

- Label each of n^k vertices with a unique k-vector of types in $\{1, \dots, n\}$.

- Connect node $\vec{u} = (u_i)_{i=1}^k$ to node $\vec{v} = (v_i)_{i=1}^k$ independently of any other connections, with probability given by the inner product of their type-to-type densities a_{u_i,v_i}:

$$\Pr\left(\vec{u}\vec{v} \in E_A^{(k)}\right) = \min[1, \langle \vec{u}, \vec{v} \rangle_A] = \min\left[1, \prod_{i=1}^{k} a_{u_i,v_i}\right]. \tag{2.5}$$

While the regularity of this type structure eases closed-form mathematical analysis [30], even enabling efficient methods for stochastically fitting the initiator matrix to an empirically-observed network [41], the model proves overly restrictive in practice for describing many real networks, on a few counts [62]:

- Inter-type densities must be identical and independent.

- While type membership ratios can be adjusted by including multiple equivalent types [40], these ratios must be identical and independent across node indices. Furthermore, the large n thus required to describe types that are small relative to the population introduces further inefficiencies; see below.

[9] The original R-MAT model normalized the $a_{\alpha,\beta}$ to sum to 1 and separately specified the total number of edges; recent models proposed by Leskovec, *et al.* instead allow the $a_{\alpha,\beta}$ to sum to an arbitrary density parameter $\bar{a}$, so that the graph densifies ($\bar{a} > 1$) or sparsifies ($\bar{a} < 1$) as it grows.

[10] In the undirected paradigm, the matrix will be symmetric, but we describe the general undirected case here.

- The population size is required to be exactly n^k, where n is the number of types and k is the number of independent node indices, with exactly one node of each combination. Not only is this regularity statistically implausible, it also severely limits both the number of indices and the ability to express types that are small relative to the population size.
- The proportional construction of the population also makes it impossible to observe dynamics *continuously*, as single nodes are added; rather, dynamic analysis [46] is limited to comparing graphs separated with sizes separated by a minimum ratio of n.

These issues may explain why, in Leskovec's original experiments in fitting Kronecker initiators [41], the fitted 2×2 initiators generally factored as

$$A \approx \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \begin{pmatrix} \alpha & \beta \end{pmatrix},$$

indicating that connection densities are modeled only by (normally-distributed) node popularity parameter, without any locality structure.

Attribute-Label Models (2009-2014)

To address concerns that iterated initiators failed to fit the locality structure of networks, a series of papers by Leskovec and various co-authors [36, 45, 49, 69] explored weakening the regularity of KrΠ's strict type-membership structure, ultimately yielding the *community-attribute graph model* (AGM) proposed by Yang and Leskovec last year [71]. In an AGM network, nodes are labeled with the *attribute* of membership (or non-membership) in each of k communities H_i. Nodes may possess an arbitrary number of such membership attributes, and are considered to fully possess each (rather than resembling some fractional mixture). Each community H_i has a coherence parameter h_i, and for each community membership that a pair of nodes share, they are linked with probability h_i.[11] Communities

[11] Probabilities compound independently, unlike Kronecker models, in which densities multiply, so in fact it is mathematically simpler to compute the marginal probability of two nodes re-

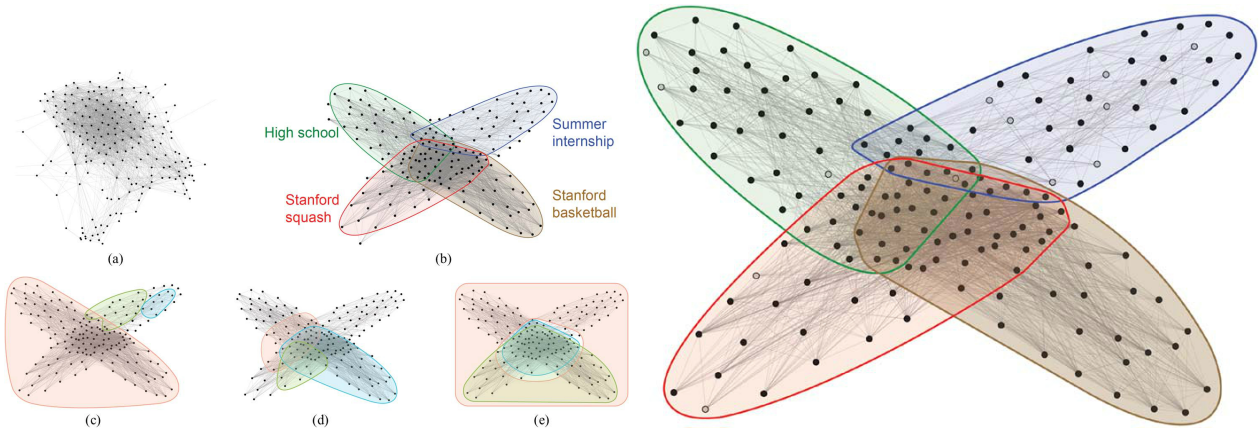

Figure 2.1.1: Figures from Yang and Leskovec [71]. **Left: (a)** a Facebook user's local friendship graph; **(b)** hand-labeled ground truth dividing it into four overlapping communities; **(c-e)** results of applying common clustering methods: **(c)** clique percolation **(d)** link clustering **(e)** mixed-memberhsip stochastic block. **Right:** the same graph with an AGM algorithmically fit by Yang and Leskovec's method.

themselves, however, are allowed to interact freely, resulting in a variety of possible intercommunity relationships beyond those possible in an (G)KrΠ model:

- disjoint (as with types at the same level of a Kronecker model);
- overlapping with independent membership probabilities (as with types at different levels of a Kronecker model);
- overlapping with (either positively or negatively) correlated membership probabilities;
- concentric, *i. e.*, $H_1 \subset H_2$.

maining *unlinked* as a product over the probabilities that they aren't linked in any community they share:

$$\Pr(uv \notin E) = \prod_{i:(u,v \in H_i)} (1 - h_i) \tag{2.6}$$

To return once more to our email-network example, consider the 'social circles' that one student might simultaneously be a part of: the mathematics department, the computer science department, the ballroom dance team, a particular class year, a particular residential house… Each community will have have its own average internal connection density, and though some may have membership distributions statistically independent from others (*e. g.*, class year is independent of house), others might not (*e. g.*, the math department has a higher concentration of ballroom dancers than average).

Furthermore, each of these communities may have additional internal structure: the executive board of the ballroom dance team is a clique in the email graph, the set of former teaching fellows in the CS department is more densely connected than department as a whole, and so on. The relationships between communities can be arbitrarily complex, allowing nested, overlapped, independent, or disjoint structures of any size and—unlike in (G)KrΠ models—substructures present in one region of the graph need not be present globally.

These possibilities, along with the freedom of scale allowed by dropping the requirement that types have integer size ratios, allow for significant expressive freedom in describing community structure. In 2014, Yang and Leskovec also presented a technique for fitting AGM structure to a network by regularized machine learning, with impressive results [71], particularly in decomposing the densely-connected core of a social network into distinct combinations of overlapping affiliations with outside groups—see Figure 2.1.1.

2.2 Why attribute labels?

The history of network structure theory can largely be characterized by the models chosen to represent 'the part of networks we care about'. As any number of ill-fated models have shown us, choice of *incorrect* models impedes our ability to accurately describe the structure of the networks under consideration, and worse, complicates the task of generalizing our insights to new samples or settings.

Thus our choice of Yang and Leskovec's community-attribute graph model

data type	models	mechanics / connection probability
(none)	ER	uniform
node degree	BA	prefers higher-degree nodes
positional	WS, CGA	prefers closer nodes (resp. in grid/tree)
node[-set]	Copy, Fire	prefers neighbors of prototypes*
label-vector	KrΠ, AGM	inner product of type-to-type densities

Figure 2.2.1: Data & mechanics for the Erdős–Rényi (ER) [24], Barabási–Albert (BA) [4, 8], Watts–Strogatz (WS) [66], community-guided attachment (CGA) [44], prototype-copying (Copy) [37], forest-fire (Fire) [44], Kronecker-product (KrΠ) [46], and community-attribute (AGM) [71] graph models. (*The mechanics of the forest-fire model modify a node's data in real time, as well.)

(AGM) [71] merits serious consideration. In particular, we wish to confirm that it is not only a good *empirical* fit to the data in practice, but that it is a good *theoretical* fit, generally abstracting the 'important parts' of a network with a reasonable degree of fidelity. It is only with this sort of theoretical validity that we will be able to interpret the differences *between* model-fittings as being structurally meaningful.

The innovation of *attribute labels* presented by the AGM appears a plausible abstraction of network structure for this purpose, though the diversity of other models that have been proposed indicates that it is not the only *prima facie* plausible candidate. Nevertheless, we argue, it is the most promising model for our data and our specific interests.

2.2.1 Forms of node data

First, we draw a distinction between the *data* of a graph model from the model's *mechanics*—the former are the relevant features of a node that predict its connection patterns; the latter is a set of (in general probabilistic) algorithms that operate on the data of the graph's nodes to generate edges.

The most common types of data used in graph models are given in Figure 2.2.1: no data, node degree, positional information, node-sets, and type-vectors. They

provide a way of categorizing graph models by their internal representation of heterogeneity between nodes and, correspondingly, the sort of structures they can or cannot naturally induce. For example, ER models are *a priori* unable to produce either fat-tailed degree distributions or locality structures, since nodes are labeled with no information that would preferentially give rise to either.

Similarly, BA over-simplifies by considering only node degree, since it is then unable to favor the production of locality structures. And while the node-sets of the prototype-copying [37] and forest-fire [44] models have been shown to induce community structure, they do it by imposing a recursive dependence on other nodes (whose own community structure was dependent on other nodes…); in practice, this iterative process proves difficult to marginalize[12] into useful conclusions about nodes on a local scale.

Positional vs. label-vector

The two remaining types of node data are *positional* (*i. e.*, encoding a location in a grid, tree, or other structure) and *label-vector* (*i. e.*, encoding one of several possible labels in each of multiple indices). Disregarding the outdated Watts–Strogatz augmented lattice model [66], we consider the remaining three graph models:

- *Community-guided attachment* places nodes in a tree structure, then connects them with probabilities based on their distance across the tree [44]. Nodes are placed into the tree in an entirely regular fasion.

- Kronecker models assign nodes a vector of labels representing types. Types are arbitrary labels in $\{1, \ldots, n\}$, where the inter-type effects are the same for every vector index and combine multicatively across indices. Labels in each index are assigned proportionally and independently.

- Community-attribute models assign nodes a vector of binary labels representing membership or nonmembership in a set of communities. Though

[12] *Marginalization*—of probabilities or other analytic measures—is the task of collapsing distributions over specific global outcomes into distributions over general categories and their local-scale effects.

model	assignment interactions	effect interactions
CGA	**regular** (tree-like)	**regular** (tree distance)
KrΠ (intra-)	**semi-regular** (disjoint; fixed ratios)	**arbitrary** (set by initiator entries)
KrΠ (inter-)	**regular** (independent proportions)	**regular** (densities multiply)
AGM	**arbitrary** (overlaps, concentricity)	**semi-regular** (coherences compound)

Figure 2.2.2: Interactions between types in both the assignment and effect of type data in the community-guided attachment (CGA) [44], Kronecker-product (KrΠ) [46], and community-attribute (AGM) [71] graph models. For KrΠ, we describe the relationships between (intra-) and across (inter-) indices.

> the labels are binary, communities are heterogeneous in both size and coherence. The joint distribution of memberships is not independent—members of one community may be arbitrarily more or less likely than the general population to be members of another.

Of the three, CGA is by far the simplest in terms of the expressivity of its community structure—nodes are identified with a single point in a hierarchical tree of communities, and nodes are considered close to each other if they share a common ancestor community with low height. Nodes are thus considered members of exactly one tier-1 community, exactly one of its child tier-2 communities, exactly one of *its* child tier-3 communities, *etc.* By contrast, label-vectors allow a node to be labeled with multiple attributes that are not themselves located in an explicit hierarchy.

While some authors have argued that some regions of the World-Wide Web can be approximated in such a hierarchical framework [44], empirical work from the Laboratory for Web Algorithmics (LAW) at the University of Milan indicates that the structures of some graphs *require* a multi-axis characterization to capture their locality structures properly [14, 65].

Later work from LAW suggested that a large category of networks, which they call *social-like*[13], are not only multi-axis in nature, but 'robustly' so [13], indi-

[13]The distinction between *social-like* and *web-like* structure is discussed later in this chapter, in §2.3.2.

	Kronecker	AGM
number of types	n	**2**
type interpretation	(arbitrary)	community (non-)membership
relationship across indices	**independent**	**arbitrary**
population construction	proportional	probabilistic

Figure 2.2.3: Comparison of Kronecker *vs.* AGM type-vector data.

cating that a multi-axis representation is necessary for our purposes.

Kronecker types versus AGM memberships

Our analysis thus suggests that flat vectors of labels will be the most useful way to express the features of nodes that affect connection structure in the social-like networks we are interested in. This leaves us with two remaining questions: *How complex should* the label system *be*? and *How regular should the relationships* between different labels *be*?

Considering the differences between Kronecker models and the AGM, we see that we face a tradeoff between:

(**Kr**Π) freedom in type structure *within* indices; regularity *across* indices	(**AGM**) binary 'type' possibilities *within* indices; arbitrary structure *across* indices.

Note here that we are considering only the dependence of *label assignment* across indices—in all cases, the *mechanics* across indices are simple multiplication/compounding. However, we might also consider a label structure that is *both* simple and regular, or one that has complex structure *and* dependence, and it is instructive to ask why we don't prefer either.

Complexity is a double-edged sword in modeling—it makes it easier to produce *descriptive* characterizations of empirically-observed structure, but more difficult to determine suitable parameter settings for *generative* uses. Nevertheless, if

a certain form of complexity is actually present in our data, our models will need some way of expressing it, but any free parameters introduced make the job of modeling more difficult, since we then need to understand how they vary across different networks.

Further consideration is thus required to determine the minimal set of free parameters our model will need to accurately describe the networks in which we are interested. Recall that fitted Kronecker initiators often fail to exhibit locality structure [41] [§2.1.3], suggesting that some freedom of inter-index interaction is required. Recent results suggest that models that include only a binary label in each index (interpreted as membership or nonmembership in a 'community') can plausibly model community structure, when they allow for complex overlap structures [45, 68, 71]. Thus we believe that a model with simple intra-index structure (membership/non-membership) but potentially very expressive inter-index relationships (dense overlap, disjointness, concentricity, *etc.*) can do a sufficient job of characterizing structural complexity in the networks we consider.

2.3 Further considerations

2.3.1 'Additive' / 'transitive' / 'competitive' community models

Besides the AGM, other modern models for mixed membership among densely-connected communities have been proposed, including clique percolation and various stochastic block models. These two, in particular, are better-established and more popular—and so deserve some discussion.

Clique percolation

Clique percolation is a combinatorial approach to community definition, which defines a community as a k-clique[14], plus any 'adjacent' k-cliques (two k-cliques are adjacent if they share $k-1$ nodes), plus any k-cliques adjacent to any of *those*

[14] A k-clique is a fully-connected subgraph of k nodes, *i. e.*, one in which all pairs of nodes are linked.

cliques, *etc.* This definition allows a natural tuning parameter for the coarseness or fineness of community detection and allows for both double and multiple overlap among communities without treating it as a special case. The structural simplicity and lack of ambiguity in the definition of 'community' makes it easy[15] to identify community structures computationally.

However, the adjacent-k-cliques model of communities is significantly limited in its expressivity by the fact that it is very nearly *transitive*. An m-transitive community model exhibits the property that, for any communities A, B,

$$|A \cap B| \geq m \Longrightarrow A = B, \tag{2.7}$$

i. e., any communities that share at least m members are the same community. Thus, '1-transitive' models have completely disjoint communities and any m-transitive model is limited to pairwise community overlaps of size strictly less than m. More generally, for a subgraph H, an H-transitive community model exhibits the property that, for any communities A, B,

$$H \subseteq (A \cap B) \Longrightarrow A = B, \tag{2.8}$$

i. e., any communities whose overlap includes the subgraph H are the same community. Thus m-transitivity is equivalent to I_m-transitivity, where I_m is the independent graph on m vertices.

The k-clique percolation community model is K_k-transitive, where K_k is the clique on k vertices. Thus, no pair of distinct communities has an overlap including a k-clique. This allows overlaps to be large, but in practice forces them to be relatively sparse. In a network exhibiting a densely-connected central core structure, as many social-like networks do [45], all but one community are severely limited in their ability to include members from the core. Figure 2.1.1 gives an example of a network with a dense core that cannot be shared between communities in

[15] all models dscussed in this section are NP-hard to resolve completely, but frequently much more tractable to approximate in practice. Clique percolation is relatively unusual in that a simple, efficient-in-practice, exact algorithm exists and is easily parallelizeable.

the k-clique percolation model. By contrast, AGM communities, which are associated with a probabilistic density property rather than combinatorial properties over subgraphs, are less constrained in their freedom to allow dense overlaps.

Stochastic block models

A stochastic block model can be thought of as a simple generalized Kronecker model, with a density matrix given by a uniform mixing (within communities) and some arbitrary intercommunity factor per community pair. Typically, membership proportions are also allowed to vary freely. Stochastic block models are therefore often used for networks where structure is dominated by variation in the density of links between some set of discrete partitions (*e. g.*, between pages paritioned by website, or employees partitioned by firm).

However, discrete stochastic block models—wherein each node is a member of exactly one community—are unable to model community overlaps naturally, requiring the overlap itself to be defined as a separate block with high density to each of its constitutents. In the case of multi-community overlaps, potentially including very few members, the proliferation of additional communities postulated becomes unwieldy.

A variant of the discrete stochastic block model, the *mixed-membership* stochastic block model, describes nodes with a mixture of fractional community memberships, with node-to-node connection likelihoods computed by product-sum over all communities they share:

$$\Pr(u \to v) = \sum_{H_i \ni u} u_{H_i} \cdot h_i \cdot v_{H_i}, \tag{2.9}$$

where u_{H_i} is u's degree of H_i-membership and h_i is the coherence parameter of H_i.

However, these overlaps are *sparser* than single-membership regions of a community, since nodes in an overlap split their connection potential among the communities in which they take part (and nodes who *share* a joint membership are *less* likely to be connected than if they had shared a sole membership in any sin-

gle community instead.) For this reason, we say that mixed-membership models have a *competitive* community structure, and have difficulty modeling dense core structures. (Again, see Figure 2.1.1 for an example.)

By contrast, community memberships in an attribute-label model such as AGM are *additive*, and the addition of a new membership serves only to increase a node's modeled connectivity with other members of the new group, rather than in the decreasing modeled connectivity with its old communities.

An advantageous result of this mechanic is that, unlike in stochastic block models, there is no need for a dedicated 'core members' group to account for the increased density in a multiply-overlapping core that is adequately described by the additive effects of the overlap. Thus, attribute-label models are more easily able to represent fine semantic structure *within* the core, including unraveling components of core nodes' popularity into affiliations with nodes located on the periphery. This capability is of significant interest for our task, allowing us to decompose bibliographic networks by topic, even in their dense core.

2.3.2 'Social-like' versus 'web-like' networks

The Laboratory for Web Algorithmics (LAW, in Italian acronym) is a group in the department of computer science at the University of Milan interested in the development and analysis of algorithms and methods for compressing and analyzing network graphs [39].

Their contributions to the study of graph structure largely stem from their focus on methods and technologies for compressing network graphs. Noting that many network graph datasets ideal for study are too large to fit in even a large computing system's working memory (a situation that has not improved over time, as graph sizes have easily kept pace with the scaling of computing hardware), they tackled the problem of *compressing* graphs, to ease their computational study. Using efficient codes that exploit commonalities in network structure, they presented a technique for compressing a large dataset drawn from the World Wide Web[16] by

[16]The WebBase dataset they used [33] contains approximately 1.2×10^8 nodes and 1.0×10^9

almost an order of magnitude [9, 10, 62].

However, the same compression techniques performed comparatively poorly on social networks (*i. e.*, an academic co-citation network, the Hollywood co-starring network, and a friendship network from the social media site LiveJournal) [13]. Later investigation into the reasons for this underperformance revealed significant differences between the structure of such social networks.

Boldi and Vigna found a distinction between *web-like* graphs, whose network structure is generally sparse and dominated by a small fraction of 'central' nodes—and *social-like* graphs, whose structure is generally dense and distributed [7]. For example, they found that the effective diameter of the Facebook social network increased by only 5% after removal of the most-central 30% of edges[17] [16]. By contrast, the same removal procedure left almost all pairs of a World Wide Web graph disconnected.

Later work by LAW demonstrated a further consequence of this resilience to disruption: the distribution of pairwise distances in social-like networks is far more regular than that of pairwise distances in web-like networks, even among networks with the same average diameter. To take Facebook as an example, a randomly-selected pair of users is separated by 3.71 intervening friends on average—but 91% of *all* pairs are separated by four or fewer. A comparable web-like graph might have ten times the variance in distances between pairs, with long distances correspondingly far more common. Since the relative variance[18] of pairwise distances (or the *spid*) is typically either well below or significantly greater than 1, Boldi, Rosa, and Vigna have suggested that the spid distinguishes two distinct classes of networks: the 'social-like' with spid $\ll 1$ and 'web-like' with spid $\gg 1$ [15].

In this work, we find ourselves primarily concerned with social-like networks,

edges [9].

[17] They explored various removal strategies, but the one upon which this result is based was to iteratively remove the most-connected nodes, along with all of their edges, until a given total fraction of edges had been removed.

[18] The *relative variance*, or *index of dispersion*, of a nonnegative distribution is the ratio of its variance to its mean. "Spid" is acrnoymic to **S**hortest **P**ath **I**ndex of **Dispersion**.

with structure driven by the contributions of many nodes, rather than an influential few. Previous work has found citation and co-authorship networks to be social-like [13], and we thus expect the structure of our networks to be driven by the distributed behavior of many nodes. Therefore, we believe that a richer model of node character (such as is provided by additive labels in an AGM) will be better able to model the network's structure, compared to a model better-suited to centralized structure, such as KrΠ or mixed-membership stochastic block.

2.3.3 Free parameters and fitting

Having chosen the AGM as our model, we then face the question: *How do we resolve the free parameters of intercommunity membership relationships?*

We can fit them to an existing graph, as in the Kronecker initiators fit by KronFit [41] or the AGMs extracted by Yang and Leskovec's machine-learning techniques [71]. However, this reliance on fitting to determine 'meta-structure' introduces limitations into the generative power of our model. Namely, it leaves us unable to scale it past the point where the fitted meta-structure can be expected to retain roughly the same form.

As a concrete example, Yang and Leskovec's fitted-AGM method can be used to produce an artificial Facebook-like network of a billion nodes by the procedure:

- Fit an AGM H_{FB} to the billion-node Facebook graph, finding both population proportions in communities (and their overlaps) and coherence parameters.
- Use the H_{FB} population proportions to randomly generate a billion nodes labeled with community memberships.
- Randomly assign links between the nodes with weights determined by community memberships.

However, if we attempt to produce a Facebook-like network of *five* billion users by simply over-populating it in the second step, then the community structure will

remain the same in shape, with every community and multi-community overlap simply five times larger.

This is not, in general, realistic—while some sorts of communities in social networks scale roughly with global population, others do not, remaining stable in size [45]. Furthermore, there is evidence that many networks *densify* and *decrease* in diameter [7, 15, 45], rather than growing apart as an overpopulated H_{FB} would. And finally, even if we only care about producing artificial networks of $\sim$1 billion nodes, we still might wish to explore the space of possible community structures that might have emerged, rather than merely considering possible populations of individuals within a single fixed structure.

Instead, we seek a generative model for the community-structure layer itself. Such a two-tiered model would allow us to provide a model not just for reconfigurations of a particular structural pattern, but a general distribution over the possible patterns that might emerge. However, there is one additional perspective we wish to consider: that of the evolution of the network structure through time.

2.4 Networks over time

Prior work regarding networks evolving over time has fallen into two distinct modes of analysis: Aggarwal and Subbian name them *maintenance methods* and *analytical evolution analysis* in their 2014 survey of the literature in evolutionary network analysis [1]. To quote from the same survey:

> - *Maintenance Methods:* In these cases, it is desirable to maintain the results of the data mining process continuously over time. For example, the results of a classification and clustering method will evolve as the structure of the graph changes over time. Therefore, the results of the methods will become stale over time, and the goal is to maintain the freshness of the end results. Correspondingly, it is desirable to provide methods that can maintain these results continuously and incrementally over time.

- *Analytical Evolution Analysis:* In these cases, it is desirable to directly quantify and understand the changes that have occurred in the underlying network. The main point to remember is that such models are focused on modeling the change, rather than correcting or adjusting for the staleness in the results of data mining algorithms on networks. Direct evolution analysis is closely related to the problem of outlier detection in temporal networks because temporal outliers are often defined as (abrupt) change points. [1]

Authors proposing 'maintenance methods' have presented evolutionary variants of conventional approaches for statistical [20] and spectral clustering [21], and stochastic-block decompositions [47]; techniques involving 'forgetting' parameters adapted for networks with bursty evolution [57, 67]; and methods designed specifically for heterogeneous, evolving networks [31]. In every case, however, their goal has been to *minimize* and *regularize* the change in network structure between successive timesteps, so as to assign the most stable labeling possible to nodes in a graph.

However, we wish instead to explore precisely how community structure evolves over time, and so align ourselves with the project of 'evolution analysis' instead. This project draws inspiration from the literature of generative network models previously discussed [3, 4, 24, 37, 38, 44, 46], to understand general node-level evolution mechanics and how they induce structural phenomena on a global scale [44]. However, to draw any meaningful conclusions from observed changes in model parameters, it is necessary to select a structural model with easily-interpretable parameters. For example, Hopcroft, *et al.* analyzed a citation network in a hierarchical clustering model, with the hope that branches of the hierarchy would correspond to academic subjects [34], with significant structural changes reflecting the emergence or consolidation of research areas in academic science. Aggarwal and Yu would later expand this work with an analysis of the expansion and contraction of communities between emergence and disappearance events as well [2].

Unfortunately, hierarchical models are known to be a bad fit for social-like networks [13, 14], and further work such as Palla, *et al.*'s experiments with clique percolation [58] has sought to use models ever better-suited to describing the networks under consideration. It is in this tradition that we use the community-attribute graph model—which is known to be superior [71] to hierarchical models, stochastic-block models, and clique percolation in modeling communities with dense overlap—to examine the evolution of *communities*, rather than hierarchies, blocks, or cliques.

Nevertheless, we draw from conceptual advances in both maintenance methods and evolution analysis to inspire both the methodology we use to identify and quantify community continuity between timeslices [29] and the technical vocabulary we use to describe forms of continuity and discontinuity [22]. The 'snapshot method'—comparing structural models fitted to rasterizations of the graph at discrete times—is one such established method [1].

When we try to pick out anything by itself,
we find it hitched to everything else in the
Universe.

John Muir

3

Data and Methods

The fundamental challenge in graph analysis is to render combinatorially-complex data regarding graph structure into concise, comprehensible forms. To analyze the evolution of community structure among authors in a bibliographic network, for example, we reduce a corpus of papers to a sequence of graph snapshots in time (*e. g.*, of author citations up to given dates), reduce each graph to a fitted AGM, identify continuities and discontinuities 'in essence' along the sequence of AGMs, and finally abstract the complex, interwoven histories of communities interacting over time into a simple set of structural-dynamic patterns.

In so doing, we discard much information by necessity and ignore a great deal of nuance. This chapter outlines our methodology and catalogues the information retained or synthesized in each step of the process.

3.1 The Computer Science citation graph

3.1.1 The ACM dataset

Our dataset is a subset of the Association for Computing Machinery (ACM)'s Digital Library database made available for academic research purposes, which we refer to hereafter as "the ACM dataset". (We thank the ACM for making the dataset available.) It encompasses 9,519 journal periodicals and 6,421 conference proceedings, a total of 337,004 published titles over a timespan of almost 63 years (1951-2014). Papers are associated with a total of 755,016 unique author IDs, though unfortunately this list includes many duplicated entries.[1] Conservative desynonymization heuristics reduce the table of authors to 550,079 author IDs. We use this set of merged authors—augmented by a few by-hand corrections in small experiments—rather than a more aggressively auto-corrected one, to avoid introducing spurious structure that would appear in later analysis.

After also recursively eliminating trees of authors with only one citation-made or cited-by record, we are left with 318,546 authors and 293,051 total papers. Considering citations and cited-by relationships together yields a total of 8,730,127 author-to-author relationships across the entire time-interval, an average of 54.8 per author. Authors appear in the dataset for an average of 12.8 years each (since they are first listed as an author on a paper), and thus add an average of 4.2 relationships per year. Note, however, that the subsamples that we chose for analysis involve much larger and faster-growing ego networks.[2]

[1] In one example, author S. Kominers appears with seven different author IDs associated with the name "Scott|Duke|Kominers", two with the name "Scott|D.|Kominers", and one each with "Scott||Duke Kominers" and "Scott Duke||Kominers".

[2] An *ego network* is a subset of a graph produced by restricting to vertices in the neighborhood of a given vertex v (and the edges between them), excluding v itself. In Appendix A, we discuss the features that make ego networks an attractive unit for subsampling graphical data.

3.2 Fitting an AGM to a static graph

Recall that an attribute-label graph model (AGM) represents network structure with a set of communities, each of which has a membership set of nodes and a coherence parameter h_i, potentially unique per-community. The probability that two nodes are connected increases the more communities they share, with each contributing community contributing a probability approximately equal to its coherence, up to a point of diminishing returns.

So, to fit an AGM to a given graph, we seek a configuration of the model parameters—membership sets $\{H_i\}_i^k$ and coherence parameters $\{h_i\}_i^k$—with respect to which the (log-)likelihood of the data is maximized:

$$\begin{aligned}\underset{\left\{\vec{\mathcal{H}}s=(H_i,h_i)\right\}_i^k}{\arg\max} \log L(\mathcal{H}s) &= \left(\sum_{uv\in E} \log \Pr_{\vec{\mathcal{H}}s}(uv \in E)\right) + \left(\sum_{uv\notin E} \log \Pr_{\vec{\mathcal{H}}s}(uv \notin E)\right) \\ &= \left(\sum_{uv\in E} \log\left(1 - \prod_{i:(u,v\in H_i)} (1-h_i)\right)\right) + \left(\sum_{\substack{uv\notin E \\ i:(u,v\in H_i)}} \log(1-h_i)\right)\end{aligned} \tag{3.1}$$

To extract a fitted AGM from a graph, it is thus necessary to simultaneously fit both the membership sets and the coherence parameters. Additionally, if the desired number of communities is not known *a priori*, then an additional fitting procedure is needed to adjust the number of communities, under some regularization constraints.

We use Yang and Leskovec's AGMFit software, in the SNAP library [42], which uses the Metropolis-Hastings algorithm [56] to perform stochastic gradient descent over the space of possible community assignments [68]. At each stage, nodes are permitted to leave a single community, join a single community, or simultaneously leave one community and join one other; the best choice from a randomly-generated sample is applied, and the process repeats. Community coherence parameters are updated by EM alternating-stage optimization, and the

number of communities itself is fit by periodically removing communities from an initially-large candidate set when they fail to contribute significantly to the l_1-regularized log-likelihood [70]. The AGMFIT algorithm terminates heuristically, typically running in time quadratic in the number of nodes in a network [68][3].

Further details about the parameters we used are enclosed in Appendix A, and we evaluate the results of AGMFIT on our data in the next chapter.

3.3 MAPPING COMMUNITY CONTINUITIES OVER TIME

As detected communities have no natural ordering or identification apart from their membership sets and coherence parameters, we post-processed the detected communities to identify continuity relationships between successive snapshots (*i. e.*, correspondence between a community detected at time t a largely similar community detected at time $t+1$, differing only by a small fraction of members who joined or left in the interval between). To accomplish this task, we considered communities at time t which had relatively large overlap with corresponding communities at time $t+1$.

However, the size of overlap between two communities A, B can be viewed from two perspectives: its size relative to A, and its size relative to B. While the ratios are equal in the case that A and B are the same size, the *influence of A on B* will be different from the *influence of B on A* if the two communities differ in size. Formally:

$$\langle A \rangle_B := \frac{|A \cap B|}{|A|} \qquad\qquad \langle B \rangle_A := \frac{|A \cap B|}{|B|} \tag{3.2}$$

$$|A| > |B| \implies \langle A \rangle_B < \langle B \rangle_A \tag{3.3}$$

where $\langle A \rangle_B$ denotes the influence of B on A and $\langle B \rangle_A$ denotes the influence of A on

[3] This result was generally reflected in our experience running AGMFIT on our samples, though we do not present quantitative data on the matter here.

B. When A and B are communities from different points in time, we refer instead to the fraction of nodes in the predecessor A_t which remain in the successor A_{t+1} (the 'influence of the predecessor on the successor') as the *survival ratio* $\langle A_t \rangle_{A_{t+1}}$ and the fraction of nodes in the successor which came from the predecessor (the 'influence of the successor on the predecessor') as the *inheritance ratio* $\langle A_{t+1} \rangle_{A_t}$, where, again, the survival ratio will be greater (resp. less) than the inheritance ratio if the community size grew (resp. shrank) over the intervening interval. If we require instead a symmetrized statistic, we may also consider the geometric mean of the two, the *(symmetric) agreement ratio*:

$$\langle A_t | A_{t+1} \rangle := \frac{|A_t \cap A_{t+1}|}{\sqrt{|A_t| \cdot |A_{t+1}|}} \tag{3.4}$$

Figure 3.3.1 gives an example of how influence, survival, and inheritance ratios can vary between differently-sized communities in practice.

3.4 Characterizing intercommunity relationships

Asymmetric influence ratios are also useful to quantify many other types of intercommunity relationships, beyond those of a community to past or future versions of itself:

- the relationship between two distinct communities within the same timeslice
- the makeup of the population of a newly-emerged community, in terms of its inheritance from other extant communities
- the dynamics of events where one community 'splits' into two or more new communities.

Any of these tasks can be conceptualized in terms of an *influence matrix*, closely related to a covariance matrix in statistics. Given a vector $(A_i)_i$ of communities (either from the same year, or from different years), each entry $M_{i,j}$ of the

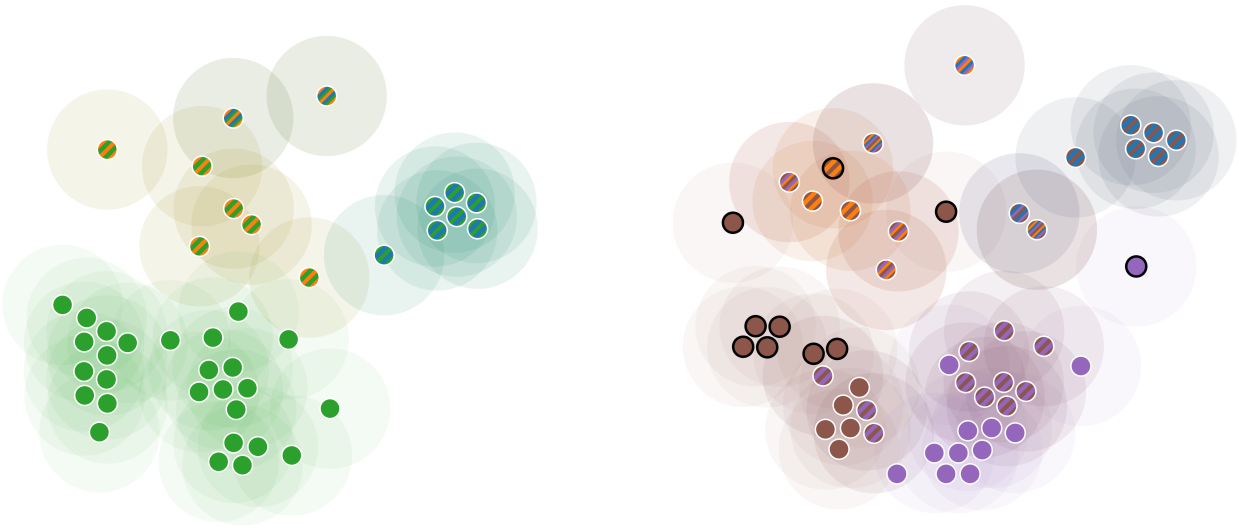

Figure 3.3.1: Left: Detail of a single community in M. Seltzer's 1998 ego network. Striped colors indicate joint membership in other communities not fully shown. Links are not shown. **Right:** The same detail in the next year, after the green community has split into two—brown and purple—with new additions circled in black.

When green split, purple inherited almost exclusively from former-green, whereas brown adopted the same absolute number of former green members, *plus* a number of new members. Thus, green has survival ratios of about $\langle G \rangle_B \approx \langle G \rangle_P \approx 70\%$ with respect to both purple and brown, but purple has a $\langle P \rangle_G = 94\%$ inheritance ratio, whereas brown has a lower $\langle B \rangle_G = 77\%$. Similarly, the agreement ratios of purple and brown with respect to green are $\langle G|P \rangle = 81\%$ and $\langle G|P \rangle = 74\%$, respectively. Note that the inheritance ratios are larger than the respective survival ratios because purple and brown are both smaller than green.

$\langle A \rangle_B$	B	1998			1999			
A		B	Y	G	B	Y	G	R
1998	B	—	0.45	0.23	**1.00**	0.47	0.20	0.07
	Y	0.44	—	0.20	0.46	**0.90**	0.17	0.17
	G	0.21	0.19	—	0.24	0.19	**0.71**	**0.69**
1999	B	**0.98**	0.46	0.24	—	0.49	0.22	0.10
	Y	0.47	**0.93**	0.20	**0.50**	—	0.20	0.15
	G	0.21	0.18	**0.77**	0.23	0.21	—	0.44
	R	0.10	0.23	**0.94**	0.13	0.19	**0.55**	—

Figure 3.4.1: 1998/1999 influence matrix between three communities in 1998 and four communities in 1999 in M. Seltzer's ego network. The upper-left and lower-right quadrants are within-year influence matrices; the upper-right and lower-left hold $\langle A_t \rangle_{B_{t+1}}$ (survival) and $\langle A_{t+1} \rangle_{B_t}$ (inheritance) matrices, respectively. Values greater than 0.3 have been highlighted, and values greater than 0.5 bolded, for emphasis.

matrix holds $\langle A_i \rangle_{A_j}$, the influence of A_j on A_i. An example is given in Figure 3.4.1, produced from the graphs visualized in Figure 3.4.2

3.4.1 The structural character of community continuity

Examination of 1639 instances of potential community continuity across 196 distinct snapshot-pairs collected from AGM fittings of twelve different ego networks[4] indicated that 69% of potential continuity instances exhibited survival ratios above

[4]Here, we examined the ego networks of Thomas Anderson, Brian Bershad, Gregory Ganger, Garth Gibson, Robert Harper, Scott Kominers, Butler Lampson, Michael Mitzenmacher, Michael Rabin, Mendel Rosenblum, Margo Seltzer, and Yaron Singer—an arbitrarily-chosen group of relatively prolific and well-connected authors in a variety of fields. On average, per snapshot, their respective ego networks had 362 authors, 8.3 communities, and 579 total memberships among 345 authors with at least one membership. An average of 127 authors per snapshot were members of more than one community, with an average of 2.85 memberships each. The largest community in each snapshot had 85 members on average, encompassing 23% of the population; the second, 78 members and 22%.

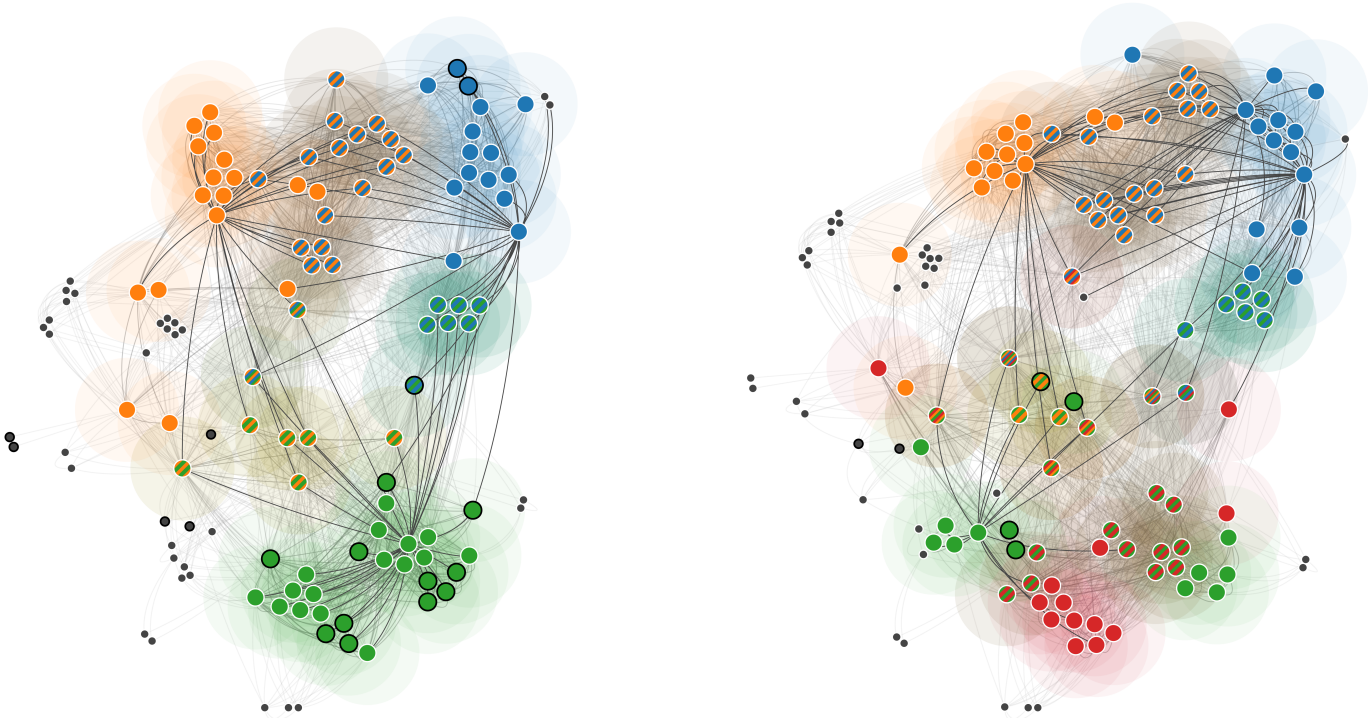

Figure 3.4.2: M. Seltzer's 1998 **(left)** and 1999 **(right)** ego networks, colored with community assignments produced by AGMFit. Striped colors indicate joint membership in two or more communities; new nodes are circled in black.

Note, as in the prior detail of this event [Fig. 3.3.1], that the 1998 green community splits into two 'sibling' communities in 1999, one (green; purple in previous figure) exhibiting significant external influence, and the other (red; brown in previous figure) overlapping very little with any communities besides its sibling.

0.8 to their closest possible successor. In 99.5% of such cases, the latter community (in the succeeding snapshot) inherited more from the former community (in the preceding snapshot) than it did from any other potential predecessor. We call such a case a *perfect match*.

In the other 0.5% of cases, *two* communities in the succeeding snapshot each inherited more from the predecessor community than from any other potential predecessor. (We call such a case a *dual match*.) Given that our networks are growing in both nodes and communities, it is inevitable that such dual matches will exist—indeed, since we will later be interested in studying them in their own right, we wish to fix a few simple criteria that will allow us to distinguish 'good' perfect matches from 'good' dual matches algorithmically.

While it would be possible for us to simply accept all perfect matches as valid, it is not clear that doing so will yield desired results. For example, there exist degenerate examples of 'perfect' matches where nevertheless the best-matched successor and second-best-matched successor both exhibit a survival ratio above 0.5—and more generally, in the presence of large overlaps between communities, the discontinuous condition of 'highest inheritance ratio' can be susceptible to noise. Thus, we prefer to treat 'perfect matches' as noisy training data to which we wish to fit a model.

However, the choice of *what* model to fit is not an obvious one. Without a generative model for community structure—as such a model is, after all, our original desideratum—we wish to use as simple a model as possible for this filter, to avoid propagating biases into our later analysis.

With knowledge of the factors we expected to be useful, we chose an extremely simple two-factor classification model:

$$\textbf{accept}\left[A_t \to A_{t+1}^{(1)}\right] \textbf{ iff: } \left(A_t = [A_{t+1}]_t^{(1)}\right) \textbf{ and } \left(\left(\langle A_t \rangle_{[A]_{t+1}^{(1)}} > c\right) \textbf{ or } \left(\langle A_t \rangle_{[A]_{t+1}^{(2)}} < c'\right)\right), \tag{3.5}$$

where $[A]_{t+1}^{(k)}$ is the community in snapshot $t+1$ to which A_t has its kth-highest

	TP	FP	FP$^{\sim}$	Recall	Precision	Precision$^{\sim}$
0.95	325	1	4	29.3%	99.7%	68.3%
0.9	516	3	7	46.5%	99.4%	77.4%
0.85	640	7	25	57.7%	98.9%	80.9%
0.8	740	25	51	66.7%	96.7%	83.1%
0.75	**848**	**51**	**82**	**76.4%**	**94.3%**	**84.9%**
0.7	928	82	118	83.6%	91.9%	86.0%
0.65	1007	118	139	90.7%	89.5%	87.0%

Figure 3.4.3: Precision and recall scores for various cutoff levels for ⟨survival ratio to best-matched successor⟩. We considered both false positives above the cutoff itself (FP) and 'close calls' (FP$^{\sim}$) within a safety margin of 0.05 points of misclassification. Ultimately, a cutoff of 0.75 was chosen manually for use, as a tradeoff between recall and precision.

survival ratio, and $\langle A_t \rangle_{B_{t+1}}$ represents the survival ratio from A_t to B_{t+1}. That is, to match A_t with its best-matched successor, we require that the successor name A_t as the predecessor it inherits the most from, *and* that either **(1)** the survival ratio is better than the cutoff c, or **(2)** no other option does better than the no-alternative threshold c'.

For our thresholds, we chose $c = 0.85$ and $c' = 0.5$ manually, after inspection of the recall and precision scores[5] they produced. (See Figures 3.4.3 and 3.4.4.) In the following chapter, we will investigate how communities in our network meet this continuity criterion over multiple snapshots, and the evolution they undergo as they do.

3.4.2 The joint character of sibling-pair inheritance

Having identified community continuities, we turn next to community *dis*continuities. We consider any event that **(1)** fails to satisfy the continuity criterion, and **(2)** where multiple communities each inherit most from a single predecessor as

[5] *Recall* is the fraction of the desired cases a classification successfully accepts; *precision* is the fraction of accepts that are correct. The trivially accepting model has perfect recall but unacceptably low precision; a model which accepts a single sample with high certainty may have near-perfect precision but low recall.

	TP	FP	FP$^{\sim}$	Recall	Precision	Precision$^{\sim}$
0	848	51	51	76.4%	94.3%	94.3%
0.3	859	51	51	77.4%	94.4%	94.4%
0.35	869	51	51	78.3%	94.5%	94.5%
0.4	904	51	54	81.4%	94.7%	94.4%
0.45	952	54	58	85.8%	94.6%	94.3%
0.5	**1009**	**58**	**88**	**90.9%**	**94.6%**	**92.0%**
0.55	1054	88	117	95.0%	92.3%	90.0%
0.6	1078	114	136	97.1%	90.4%	88.8%
0.65	1107	136	150	99.7%	89.1%	88.1%

Figure 3.4.4: Precision and recall scores for various 'no-alternative thresholds', which pass cases failed by the above test when ⟨survival ratio to second-best-matched successor⟩ is *below* the no-alternative level. We considered both false positives below the threshold itself (FP) and 'close calls' (FP$^{\sim}$) within a safety margin of 0.05 points of misclassification. Ultimately, a cutoff of 0.5 was chosen manually for use, as a tradeoff between recall and precision.

a *splitting event*, taking any predecessor–successor relationship with a survival ratio above 0.5 as a *parent–child* relationship. (The relationship between children is, naturally, *siblings*.)

Having thus identified community-splitting events on a structural basis, we can now examine the structural character of 'new' communities at the time that they are first identified by AGMFit (either as *de novo* arrivals, or as children in a splitting event). In neither case do we expect this 'appearance' to correspond to a discrete event in the bibliographic universe; rather, it occurs at the point in time that modeling the structure becomes 'worth it' with respect to our regularization heuristics. Thus, we consider 'new' communities as generally representative of nascent proto-communities in general, even before the point at which they are observable algorithmically.

We are particularly interested in the relationships between newly-emerging communities, their closest 'siblings', and the other communities with which they respectively overlap. In Figure 3.3.1, for example, we observe one community in 1998 'splitting' into two communities in 1999. (Since neither has a survival ratio

above 0.75, our criterion does not consider either to be the continuation of the parent—rather, we treat both as simultaneous new appearances.)

Both exhibit relatively high survival ratios with respect to their parent, but one (red) inherits almost exclusively from the parent itself, whereas the other (green) both inherits the bulk of the parent's external influence and acquires significant additional influence from previously-unaffiliated nodes. Furthermore, we see that their *intersection*, considered alone, inherits a disproportionate amount of the parent's external influence.

We might ask, then, how common these patterns are; more generally, we might ask how the communities resulting from a split differ from each other and from their mutual intersection—in terms of the external structure they inherit—in a 'typical' splitting event.

I am a part of all that I have met.

Alfred Tennyson

4

Experimental Results

For ease of reference, plotted figures are collected in §4.5, at the end of this chapter.

4.1 Preliminaries

As discussed in the previous chapter [§3.4.1], we apply a simple criterion on the survival and inheritance ratios between communities to distinguish continuity and splitting events from non-events, as well as from each other. From our 1639 instances of potential community continuity discussed previously [§3.4.1], we find 146 qualify as *binary splitting events* (*i. e.*, splitting events which produce exactly two child communities), out of a total of 336 total potential splits. Figure 4.5.1 shows that these events take place in snapshots of a wide range of sizes—from snapshots of fewer than a hundred nodes total, to snapshots including communities of a few hundred nodes. Similarly, the snapshots from which we draw samples include be-

Term	Subset	Size (% / Parent)
Parent	A	100%
Child/Sibling	B ; C	30%
Intersect	$B \cap C$	33%
Wing	$(B \cap \neg C) \cap A$; $(C \cap \neg B) \cap A$	31%
Extension	$(B \cap \neg C) \cap \neg A$; $(C \cap \neg B) \cap \neg A$	28%

Figure 4.2.1: Terminology for segments of a sibling pair. The final column provides rough average sizes; see Figure 4.5.3 for a more complete table.

tween 3 and 21 communities, so the effects we observe are not particular to any given network size.

Across the range of network sizes, the number of communities fit grows approximately in proportion to the number of authors in the network. (See Figure 4.5.1.) This is not a particularly profound observation about network structure *per se*, since the absolute number of communities is controled by AGMFit's regularization heuristics, but it is useful to remember as we consider samples across a range of sizes.

4.2 Anatomy of a sibling pair

Each binary splitting event involves a single parent community and two child/sibling communities. For the purposes of structural anaysis, we consider the siblings' intersection and their disjoint differences as distinct subsets. Furthermore, for each of the disjoint differences, we make a distinction between the portion inherited from the parent (the 'wing') and the portion which was adopted from elsewhere (the 'extension'), including new nodes, previously unaffiliated nodes, and nodes from other communities added to the child. These segments each exhibit distinct character in terms of their respective relations to other communities, as we see below.

These segments generally assume roughly similar proportions across a range of sizes spanning most of an order of magnitude. Figure 4.5.3 indicates that the

		Intersect	Lg. Wing	Sm. Wing	Not Inherited	Lg. Extension	Sm. Extension
parent's	μ	33.6%	38.2%	23.9%	4.3%	36.6%	20.2%
nodes	σ^2	8.7%	6.7%	6.8%	4.9%	15.7%	13.7%
parent's	μ	53.0%	22.5%	18.2%	6.3%	27.9%	19.1%
CRC	σ^2	13%	20.7%	16.7%	33.7%	12.2%	10.3%

Figure 4.2.2: Fraction of the parent's nodes and CRC inherited or adopted by each segment among the samples studied. All fractions are with respect to the parent community; the first four columns are inherited from the parent, but the last two are adopted from external sources. 'Lg.' 'Sm.' denote statistics of the larger (resp. smaller) of the two wings or extensions in each sample.

relative sizes of the head and each wing and intersection are almost nearly linear in the size of the overall union and correspondingly, that the fraction of the union that each segment represents is roughly invariant over community size. This invariance suggests that we may plausibly compare the structural dynamics of large and small splitting communities in the same terms, as they represent the same general sort of event. We will revisit this hypothesis throughout our analysis.

Figure 4.2.2 reports the average size of each segment observed in our samples; note that the five segments are roughly of equal size. As with the number of communities by author count, this fact is not itself particularly enlightening, as it is largely an artifact of the parameters we use to regularize number of communities, but it too will be useful to keep in mind throughout our analysis of the structural character of each segment.

The *community relation count* (CRC) also presented there is the number of memberships a node or set of nodes has with communities *besides* either of the siblings. Thus a node in the intersect, but with no external memberships, would have a CRC of 0, whereas a node in one of the wings with two other memberships would have a CRC of 2. The total CRC is intended here as a metric for the amount of total external influence exerted by (and on) a particular region of the split.[1]

[1]In a more sophisticated treatment, we might use a proper structural measure to quantify the importance of particular nodes. However, here CRC serves as a simple linear approximation, in

4.3 (The lack of) 'inter-sibling rivalry'

In §3.4.2, we asked about the relationship between siblings, particularly with respect to the degree to which their inheritance patterns were correlated or anticorrelated. In the particular, we observed an instance where one sibling both inherited and adopted significant external influence, whereas the other remained almost entirely limited to its parent's former membership, with relatively few overlaps with other communities besides those mediated by its more-'social' sibling.

This pattern, however, is not well-supported in the data on sibling-pairs—we observe no such modality on the distribution of 'sociality' across our sample of binary sibling pairs. Consider Figure 4.5.4, which indicates that bimodality is not generally supported in the data. While this result may appear surprising, it serves to highlight that the term splitting *events* is something of a misnomer—in actuality, we are observing nascent community structure being brought to light, not an significant shift in the global community structure.

4.4 Per-segment inheritance dynamics

We consider, then, what the emergence of nascent structure *can* tell us: the degree to which each of the segments is generally connected to external communities before and after the split. Figure 4.5.5 reports the fraction of the parent's members with external relations which is inherited by each segment, respectively. In all cases, observe the roughly linear relation on a linear-logarithmic scale, suggesting a power-law distribution, as is present in general throughout the network. Note also that:

- Considering only counterparts to whom the parent was *weakly* connected, the intersect is approximately three times more likely than any other single segment to inherit the parent's high-CRC members. In total, it inherits about 60% of the high-CRC members in this group, and each extension adds an additional 20% to the 20% that its wing inherits.

the absence of a model-driven approach.

- However, among communities to whom the parent was *strongly* connected, the intersect is overwhelmingly more likely to inherit high-CRC members. In the sample observed, the intersect was approximately 50 *times* more likely to inherit a member from the parent with a CRC of 4 to strongly-related communities.

Ultimately, the wings end up with only slightly fewer high-CRC members than the intersect, as one might expect for two randomly-selected communities with comparable overlap size [Figure 4.5.7]. However, the vast majority of the new high-CRC members in the wings or extensions (*i. e.*, outside of the intersect) are from *new* counterparts, to whom the parent community was not previously connected.

4.5 Figures

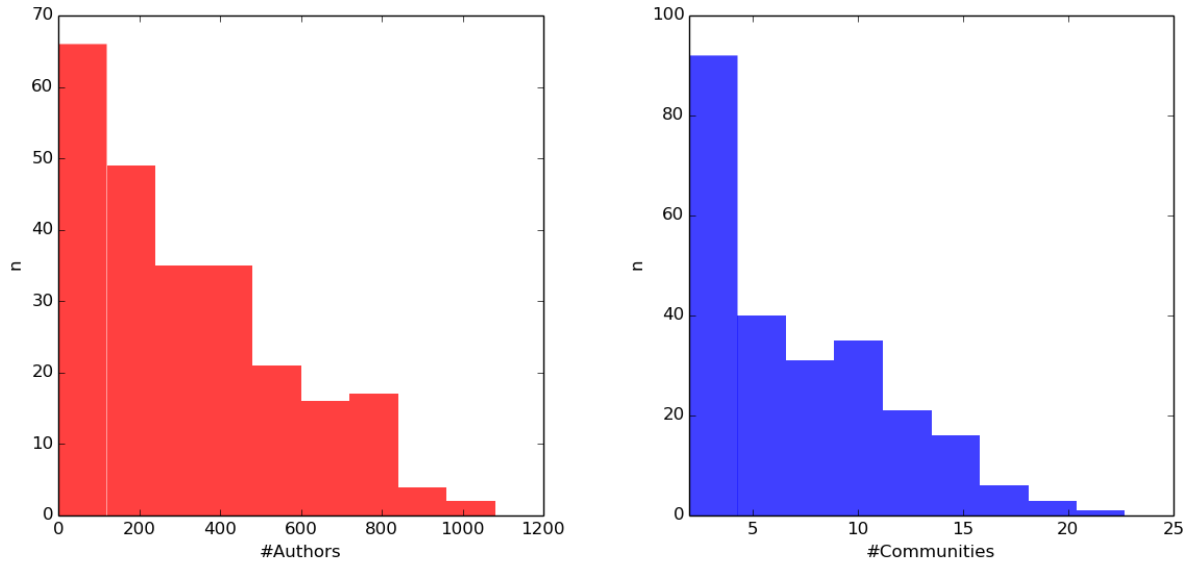

Figure 4.5.1: Samples by number of authors and number of communities fitted.

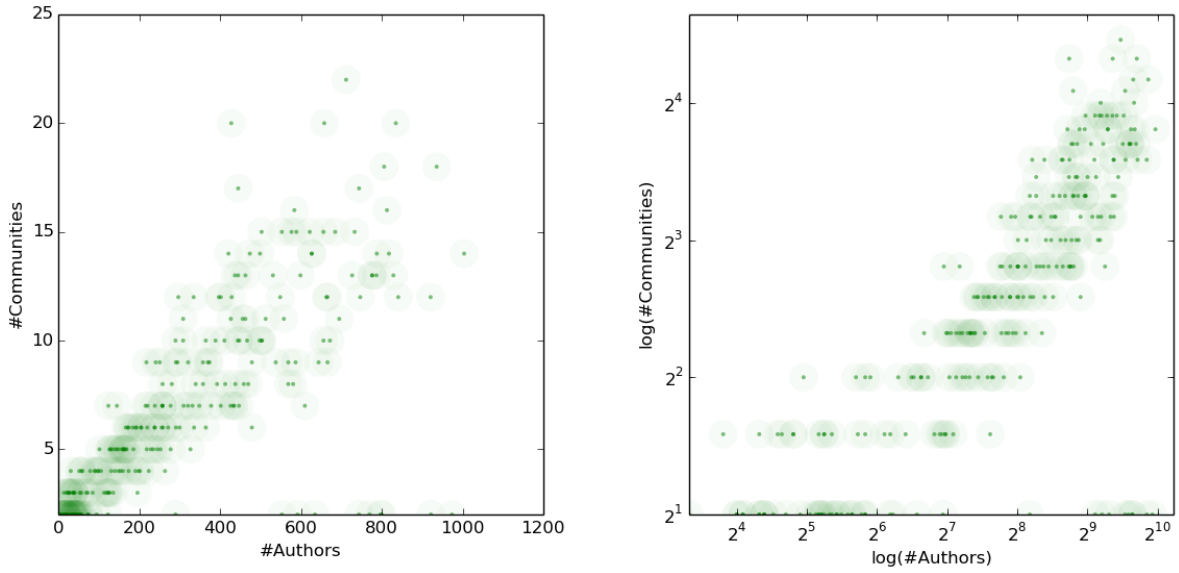

Figure 4.5.2: Number of authors *vs.* number of communities fit

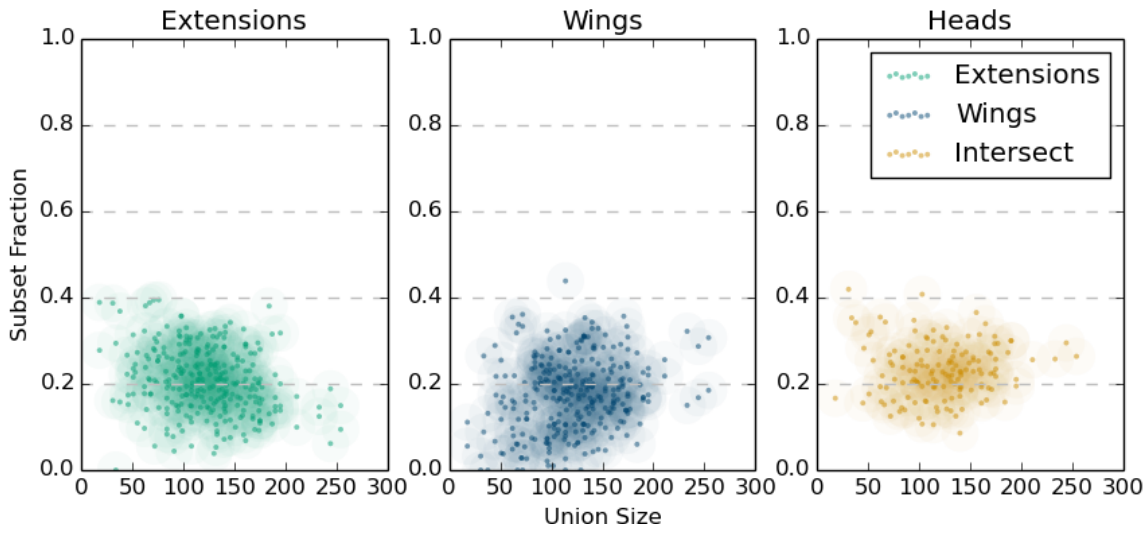

Figure 4.5.3: Total size of a splitting community *vs.* the proportions of the intersection ('head'), the inherited disjoint components ('wings'), and the adopted disjoint components (the 'extensions'). The fraction of the union that each segment makes up is roughly independent of union size; the trends are flat with $p = 0.007, 1.4 \times 10^{-5}, 0.02$, respectively.)

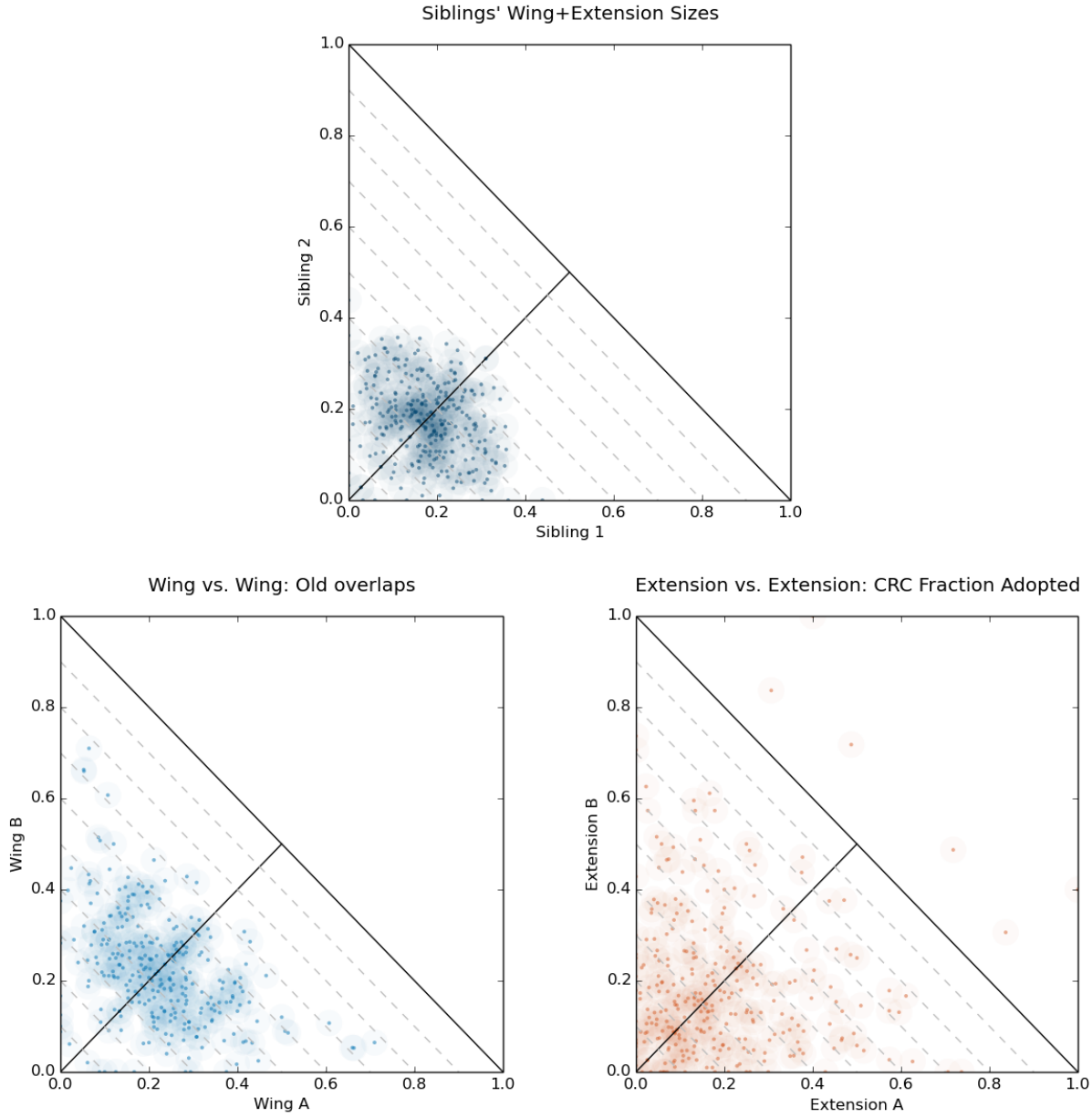

Figure 4.5.4: Top: Siblings' respective total sizes (excluding the intersect). **Bottom left**: Siblings' respective wing sizes. **Bottom right**: Siblings' respective extension sizes. Note in all cases that distributions are unimodal, indicating that the extreme difference of character observed in §3.4.2 is atypical.

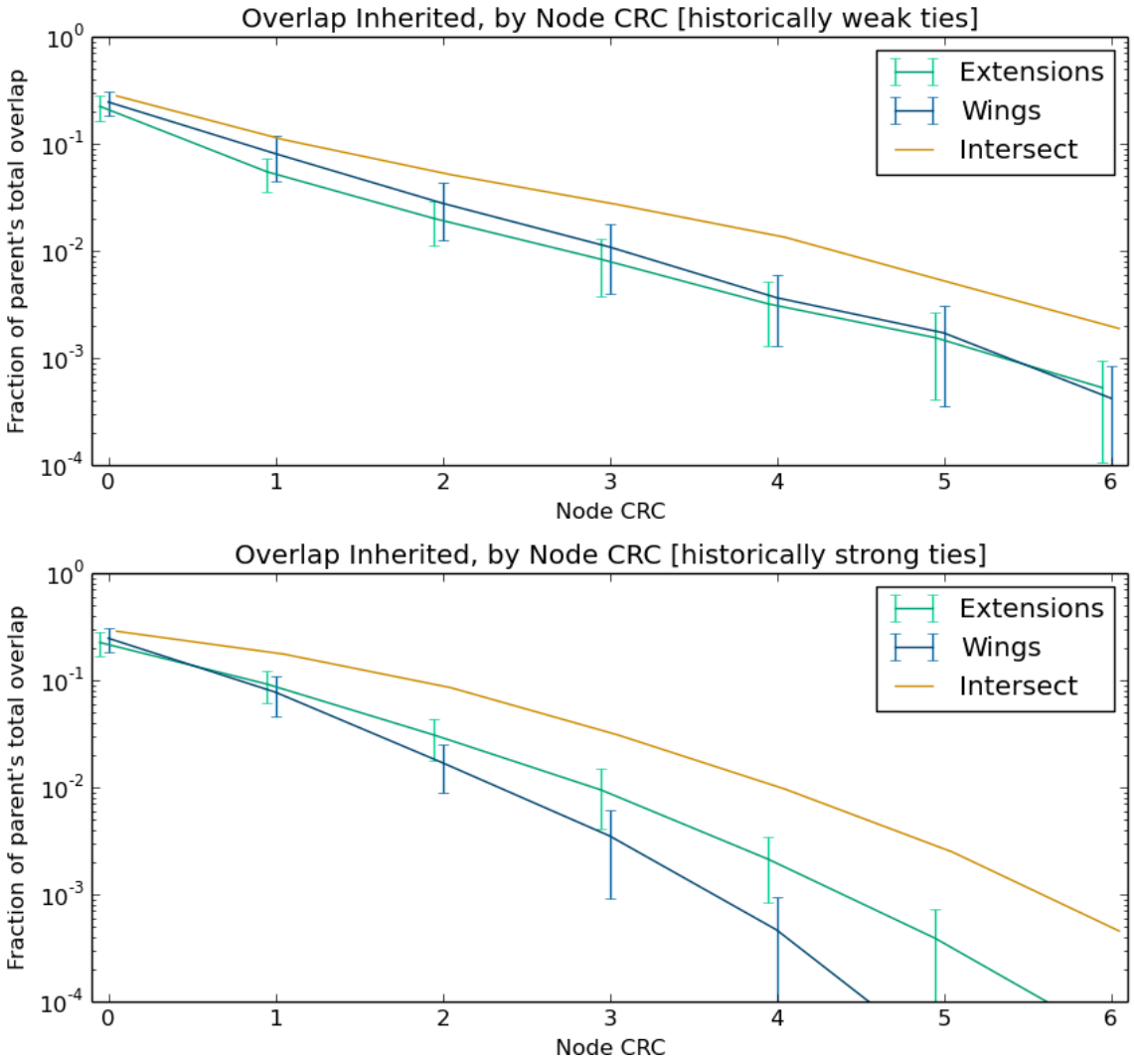

Figure 4.5.5: Among the 71 splitting events with > 125 final members, proportion of externally-connected members inherited by the extensions, wings, and intersect, respectively. The horizontal axis limits consideration to nodes with a given community relation count or higher. Vertical bars on the wings and extensions lines indicate the average spread between the sibling of a pair that inherited more and the one who inherited less.

Top: with respect to external counterparts with whom the parent community was *not* strongly connected. **Bottom**: with respect to external counterparts with whom the parent community *was* strongly connected. Note the vertical log scale.

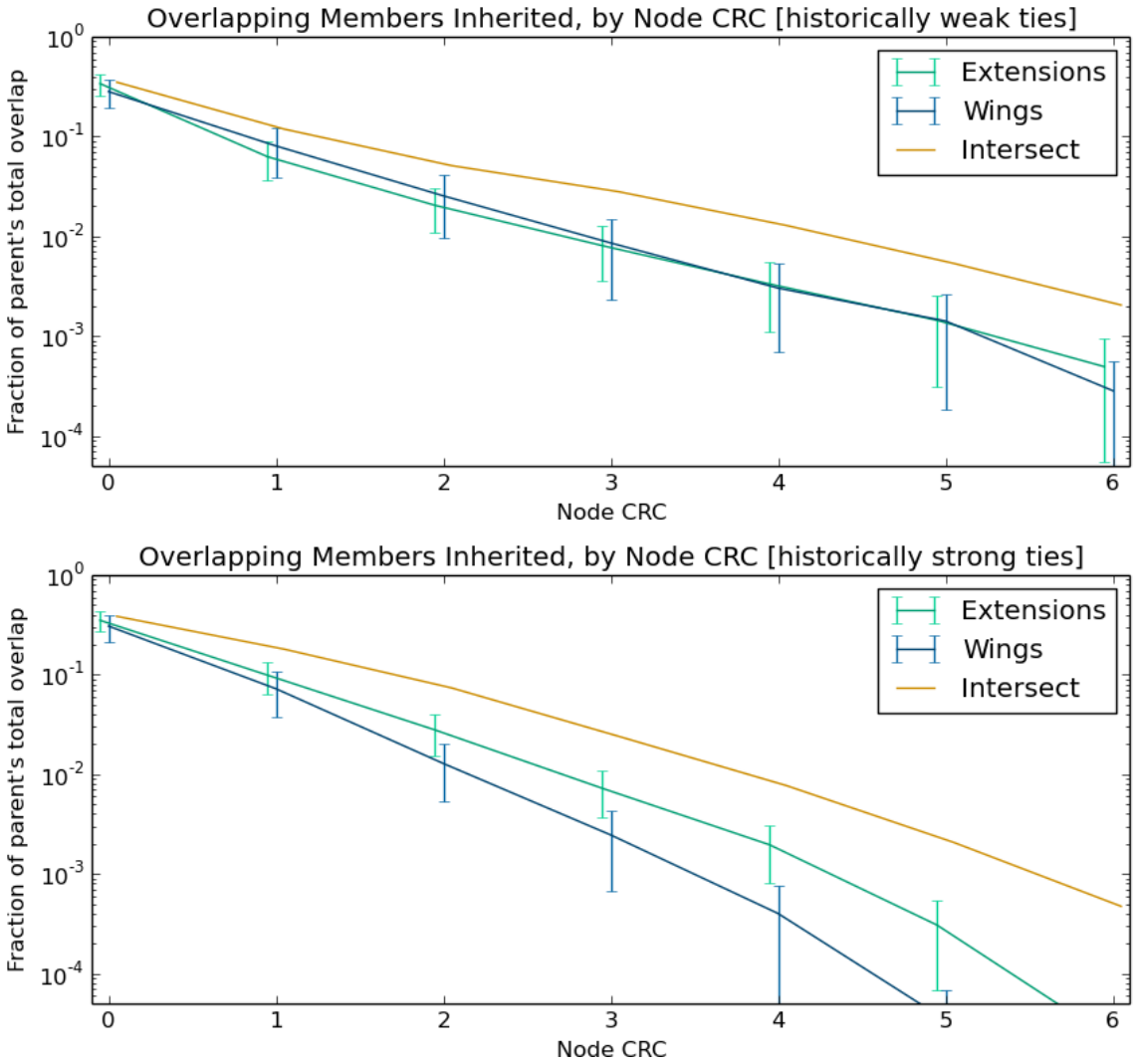

Figure 4.5.6: Same as preceding figure [4.5.5], but considering all 146 splitting events: proportion of externally-connected members adopted by the extensions, wings, and intersect, respectively.

Note in both this and the preceding figure that the curves in the top chart remain relatively close—indicating that influence from communities weakly tied to the parent is inherited almost evenly by all parts of the sibling pair. By contrast, the curves in the bottom chart diverge, indicating that influence from communities strongly tied to the parent is inherited disproportionately by the siblings' intersect—especially influence in the form of high-CRC nodes.

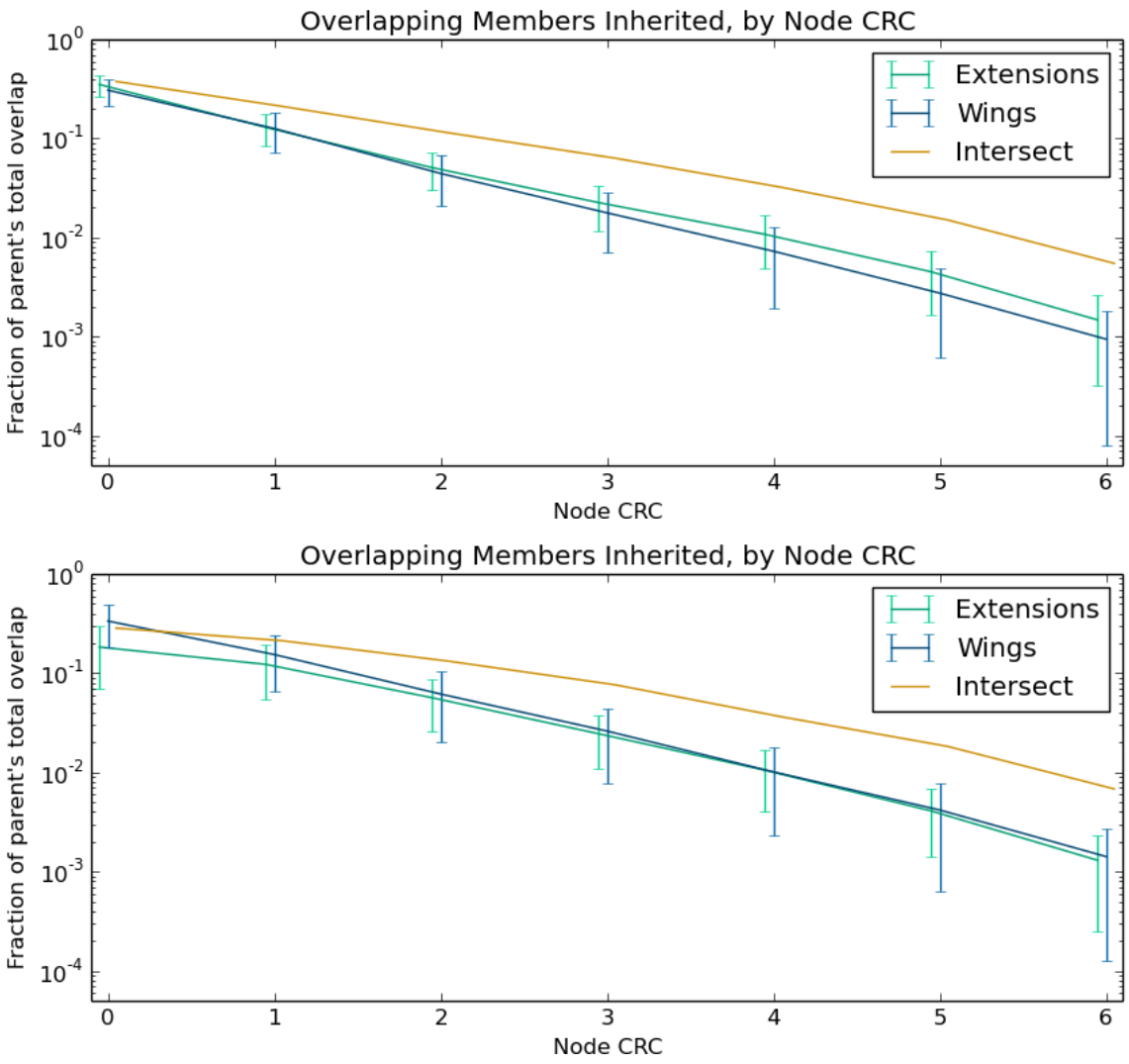

Figure 4.5.7: General high-CRC adoption patterns (disregarding old/new and intercommunity tie strength) are largely similar between sibling pairs **(top)** and comparable pairs of a new child and a non-sibling community with which it shares a similar overlap **(bottom)**.

[The computer scientist] builds her castles in the air, from air, creating by exertion of the imagination. Few media of creation are so flexible, so easy to polish and rework, so readily capable of realizing grand conceptual structures…

Fred P. Brooks

5 Concluding Notes

5.1 Contributions

We presented CAMBRIDGE, a general technique for evolutionary analysis of a dynamic network via a community-attribute graph model (AGM). By making comparisons between AGMs fit algorithmically to each of a series of annual snapshots of a graph, we are able to quantify the evolution of graph structure in terms of the relationships between communities over time. Applying our methodology to a bibliographic network of more than a quarter-million nodes and eight million edges, we presented evidence that, in the process of community division, well-connected nodes that are part of strong intercommunity ties behave differently than equally well-connected nodes that are part of weak intercommunity ties.

Previous work by Leskovec *et al.* has established that the AGM model is superior to other models traditionally used for evolutionary analysis, particularly in

its ability to meaningfully describe the dense core of a network in terms of nodes' relations to the periphery structure, and for this reason, we believe that our attribute-label evolutionary analysis is uniquely able to reveal insights about the evolution *of intercommunity interaction*, rather than the evolution of communities considered individually.

Our approach is not tailored to any particular network, or even any specific data domain. Rather, it is intentionally general, seeking to encompass and describe social interactions broadly characterized, and thus to ground the study of a diverse class of social-like networks in common principles robust and flexible enough to support the weight of future theory.

Finally, our work suggests many promising directions for future research.

5.2 Future directions

5.2.1 Model-based continuity criteria

We have only just begun to use the power of attribute-label models in evolutionary analysis, particularly for informing model assignments using the historical information that is available. Maintenance models such as FacetNet have demonstrated the utility of considering the problems of static model fitting and intertemporal continuity detection in one unified step, rather than as separate problems [47], and we believe that attribute labels will prove particularly amenable to this sort of maintenance fitting.

5.2.2 Dynamic-aware attribute labels

Our results concerning the structure of recently-split communities indicate that there is latent structure in dynamic social-like networks which is not completely captured by static attribute-label models. While all models ultimately face this limitation in the process of abstraction and generalization, theoretical advances in evolutionary analysis may be able to inform useful extensions of structural models to provide 'first-order corrections'.

For attribute-label models, we imagine that these could either take the form of *history-aware labels*—community labels whose interactions are influenced by their individual and joint histories—or *history-aware nodes*, whose relations with their own community labels (as well as those of others) is influenced by their personal histories. Our study of the history of network models indicates that managing the tradeoff between over-abstraction and over-complexification will be crucial.

5.2.3 Generalized dynamic tensor models

In one sense (and especially for maintenance models), the power afforded by considering a network in time is that it gives us the ability to view the same object 'from multiple perspectives'—just as the rising waters of a flood can show more about the contours of the land than can any static water level. Recently, various authors have proposed *tensor methods* (which, broadly speaking, treat graphs parametrized by implicit inputs) for network analysis [6]. For an additional axis along which to consider a network, some tensor methods use an external, natural phenomenon—such as time—while others use artificial or virtual phenomena—*i. e.*, by resampling graph data while 'controlling' for the presence of various nodes [63]. We expect that the dynamic and combinatorial nature of tensor methods may be particularly amenable to the techniques we have explored in dynamic attribute-label analysis, particularly in providing efficient fitting methods and natural evolutionary models.

We're expert enough to change the laws of physics
temporarily; how hard can wiring be?

Carl J. Romeo

A Technical Detail

This appendix lays out the most important implementation details of the analytic tools we developed in connection with this research.

A.1 Network construction

We explore networks by projecting them into graphs, ideally in ways that reflect general features of the network, rather than artifacts of the projection.

A.1.1 Undirected, Unweighted Citations

Many different graphs may be extracted from a full bibliographic dataset—co-authorship, co-citation, mutual citation, directed citations, citation counts (either directed or undirected) *& c.*—but in this work, we consider only undirected graphs of unweighted citations (where two authors are connected if at least one has cited the

other in a published work within our dataset).

We expect many of our findings to extend generally to other forms of bibliographic data; our focus on a single network type was intended only to facilitate exploratory analysis in depth. Future work might explore the commonalities and differences in applying our time-dynamic-community analysis to other bibliometric networks, bibliography of other academic fields, and to other settings where the time-dynamic nature of network structure is of importance.

A.1.2 Recursively Trimmed Networks

All networks analyzed in this work have been pre-processed by removing disconnected authors and trees of authors from the graph. In the latter case, this is accomplished by recursively removing "leaves": any author who, in the time-interval under consideration, cited or was cited by at most a single colleague.

In terms of community-detection, we consider their case uninteresting—either their single connection is sufficient for them to be included in one or more of the communities that their connection is a member of, or it is not—and, since we do are not analyzing popularity or other features which are might be influenced by the addition of leaves and trees, the operation does not affect the model predictions over classifications, merely simplifying the calculations involved.

A.1.3 Ego Networks

Subsample construction is a problem of particular importance—and particular difficulty—in graph analytics. In most *tabular* settings, the complexity of analysis scales linearly in the number of data entities and if size is ever prohibitive, useful subsamples can safely be drawn at random from the data.

In graphical settings, however, network dynamics make it dangerous to discard even unbiased segments of the data. It is often useful, therefore, to be able to identify a subsample which is representative on a *local* scale, at least. One such popular technique uses *ego networks*—the subgraph of nodes which are connected to an "ego" node, minus the ego node itself—as locally-representative subsamples

[32, 49]. While popularity and importance measures of individual nodes is skewed by this analysis (a node that's very close to the ego can appear central in an ego network, while being relatively unimportant on a global scale), the structure of ego networks can often shed insight on network-structural patterns observable on similar scales throughout the network. As our work concerns the trans-temporal dynamics of community structure on local scales, ego networks are well-suited to our analysis.

Elsewhere in this work, when we mention ego networks, we have constructed them by the process we described above for general networks: first, the candidate set of authors is initialized to the set of authors citing or cited by the ego (in a paper published before the considered date, for time-based samples). These authors are connected with undirected, unweighted edges representing citation relations (up until the considered date) and recursively trimmed to remove isolated nodes, leaves, and trees with a single point of attachment.

A.2 Processing

A.2.1 AGMFit

All networks were fit with AGMs by Jure and Leskovec's AGMFit software [42], described in §3.2. Except where otherwise noted, networks were fit to an ε-coherence parameter of 0.05, a number of communities regularized by AGMFit's own l_1-regularization-based heuristics, and with default termination-heuristic parameters.

Each run of AGMFit was run separately, rather than seeded with the results of the prior timeslice. (The number of communities was similarly re-fit on each run.) We made this choice to avoid biasing our analysis of trans-temporal community dynamics—it is likely that were each fitting seeded with information from the preceding timeslice, the fitted communities might exhibit trans-temporal correlation solely by virtue of this algorithmic seeding. While we expect that future attribute-label models may be designed to make use of historical network dynam-

ics in community detection, we were not prepared to make use of this information at this point.

A.2.2 Community continuity detection

After detecting communities, we post-processed the detected structure to identify continuity relationships, matching the set of communities detected in one snapshot with their best matches for "essentially the same community" in adjacent snapshots. As AGMFit reports communities identified only by their membership sets and coherence parameters, we performed this continuity detection based only on communities' respective membership vectors.

Given two communities A, B, we define their inner product $\langle A|B\rangle$ (or their *agreement ratio*) as the size of their intersection, normalized by the geometric mean of their respective sizes:

$$\langle A|B\rangle := \frac{|A \cap B|}{\sqrt{|A| \cdot |B|}}. \tag{A.1}$$

We define $\langle A\rangle_B$ the *influence of B on A* (or in some contexts the *survival ratio A to B* or the *inheritance ration of A from B*) as the size of their intersection, normalized by the size of A instead:

$$\langle A\rangle_B := \frac{|A \cap B|}{|A|}. \tag{A.2}$$

By analysis described in §3.4.1, we fixed a continuity criterion in terms of the survival ratio of a community to its most-related successors:

$$\begin{aligned}\textbf{accept } \left[A_t \to A_{t+1}^{(1)}\right] \textbf{ iff: } &\left(A_t = [A_{t+1}]_t^{(1)}\right) \textbf{ and } \\ &\left(\left(\langle A_t\rangle_{[A]_{t+1}^{(1)}} > 0.8\right) \textbf{ or } \left(\langle A_t\rangle_{[A]_{t+1}^{(2)}} < 0.5\right)\right),\end{aligned} \tag{A.3}$$

where $[A]_{t+1}^{(k)}$ is the community in snapshot $t+1$ to which A_t has its kth-highest survival ratio, and $\langle A_t\rangle_{B_{t+1}}$ represents the survival ratio from A_t to B_{t+1}. The require-

ment that successor communites may only be in relationship with the predecessor from whom they inherit most eliminates the need for a tiebreaker mechanism, so we identified successorship by direct search on community sets—which are at most a few dozen communities each in any pair of snapshots.

There is always a bigger fish.

Qui-Gon Jinn

B

Distributions in Theory

While locality was the first phenomenon of network graphs to be seriously studied [66], node-degree distribution has seen at least as much analysis, from some of the first papers in the field [3, 8]. This appendix provides a brief overview of the three families of distributions most commonly referenced and observed. In particular, we will attempt to shed insight onto why two popular distributions—the power-law and lognormal distributions—are so often informally conflated.[1]

[1] A version of this discussion previously appeared in our survey of techniques for social-like network analytics [62].

B.1 Distribution definitions

Distribution B.1.1. The *Poisson* distribution with parameter λ is proportional to $\lambda^x/x!$ in x.

It has mode λ, mean λ, and asymptotic behavior as $\widetilde{\Theta}\left(x^{\Theta(x)}\right)$.

It is a common result by Erdős that an Erdős-Rényi graph has node-degree distribution well-approximated by a Poisson distribution with parameter $\lambda = np$ [24, 54].

Distribution B.1.2. The *power-law* (or *Zipfian* [72–74], or sometimes *Pareto* [59]) distribution with parameter (or exponent) γ is proportional to $x^{-\gamma}$ in x.

It has mode 1, mean $\frac{H_{N,s-1}}{H_{N,s}}$ (where $H_{N,s}$ is the Nth generalized harmonic number, and asymptotic behavior as $\Theta(x\lambda)$.

The basic Barabási-Albert model with strictly linear attachment preference induces a node-degree distribution approaching power-law with $\gamma = 3$ [8]; affine-linear preferences with positive intercept induce a power-law with $\gamma \in (2, 3)$ [60]. Note that as $\gamma \to 2$ in a power-law node-degree distribution, the total number of edges ($\approx x^{-(\gamma-1)}$) diverges to ∞.

It is a folk law that many emprically-observed distributions in network graphs can be approximated as power-law with $\gamma \approx 2.2$ or ≈ 1.2,[2] though other values in the interval $(2, 3)$ are also observed.

[2] Note that integrating over a power-law distribution with $\gamma = 2.2$ yields a power-law distribution with $\gamma = 1.2$.

dist.	asymptotic	asymptotic-log
Poisson	$\widetilde{\Theta}\left(x^{-\Theta(x)}\right)$	$\Theta(-x \log x)$
lognormal	$\Theta\left(x^{(\mu-\log x)^2/\sigma^2}\right)$	$\Theta(-(\log x)^3)$
power-law	$\Theta\left(x^{-\lambda}\right)$	$\Theta(-1)$

Figure B.2.1: Asymptotic and asymptotic-logarithmic behavior of Poisson, lognormal, and power-law distributions.

Distribution B.1.3. The *lognormal* distribution with parameters μ, σ^2 is proportional to $\exp[-(\ln x - \mu)^2/\sigma^2]$ in x, or $N(\mu, \sigma^2)$ in $\ln x$.

It has mode $\exp[\mu - \sigma^2]$, mean $\exp[\mu + \sigma^2/2]$, and asymptotic behavior as $\Theta\left(x^{(\mu-\log x)/\sigma^2}\right)$.

B.2 Asymptotic behavior and similarities

Figure B.2.1 reports the asymptotic behavior of these three distributions, as well as their asymptotic logarithmic behavior, to illustrate the character of the asymptotic tail. (Less-negative asymptotic-logarithmic behavior implies a thicker tail.)

Note that while the distributions' tails become fatter moving down the table, the gap between the Poisson and lognormal asymptotic-logarithmic tails is polynomial, while the gap from the lognormal to the power-law is merely polylogarithmic. For this reason, it is often much more difficult to distinguish power-law distributions from lognormals than it is to distinguish Poisson distributions from either.

Mitzenmacher additionally proposes a few simple mechanisms by which power-law and lognormal distributions each arise [53], including the following result:

Theorem B.2.1. *In [53]. A mixture of lognormal distributions with identical σ^2 and μ mixed according to an exponential (e^{-rx} in x) distribution is marginally power-law.*

This procedural identity provides some insight into the similarity and distinction between power-law and lognormal distributions in generative terms, a fuller discussion of which is beyond the scope of this survey. Interested readers are directed to Mitzenmacher's survey work [52, 53] as an introduction to the topic.

In summary, though, power-law and lognormal distributions are significantly more similar in tail behavior than either is to the Poisson distribution, and for many purposes, the distinction is disregarded. Writes Mitzenmacher [53]:

> From a more pragmatic point of view, it might be reasonable to use whichever distribution makes it easier to obtain results … The recent work argues that for at least some network applications the difference in tails is not important[, but w]e believe that formalizing this idea is an important open question.

Bibliography

[1] C. Aggarwal and K. Subbian. Evolving network analysis: A survey. *ACM Comput. Surv.*, 47, Apr. 2014.

[2] C. Aggarwal and P. Yu. Online analysis of community evolution in data streams. In *Proc. SDM '05*, pages 56–67, 2005.

[3] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *Proc. ACM STOC '00*, pages 171–180, 2000.

[4] R. Albert and A.-L. Barabási. Topology of evolving networks: Local events and universality. *Phys. Rev. Lett.*, 85(24):5234–5237, Dec. 2000.

[5] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2002.

[6] F. Arabshahi, F. Huang, A. Anandkumar, C. T. Butts, and S. M. Fitshugh. Are you going to the party: depends, who else is coming? [learning hidden group dynamics via conditional latent tree models]. In *Proc. ICDM '15*, pages 697–702, Nov. 2015.

[7] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *Proc. WebSci '12*, pages 33–42, 2012.

[8] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct. 1999.

[9] P. Boldi and S. Vigna. The WebGraph framework I: Compression techniques. In *Proc. WWW '04*, pages 595–601, 2004.

[10] P. Boldi and S. Vigna. The WebGraph framework II: Codes for the World-Wide Web. In *Proc. DCC '04*, 2004.

[11] P. Boldi and S. Vigna. Four degrees of separation, really. In *Proc. ASONAM '12*, pages 1222–1227, 2012.

[12] P. Boldi, M. Santini, and S. Vigna. Permuting web graphs. In K. Avrachenkov, D. Donato, and N. Litvak, editors, *Algorithms and Models for the Web-Graph, 6th International Workshop*, volume 5427 of *Lecture Notes in Computer Science*, pages 116–126. Springer, 2009.

[13] P. Boldi, M. Santini, and S. Vigna. Permuting web and social graphs. *I'net Math.*, 6(3):257–283, 2010.

[14] P. Boldi, M. Rosa, M. Santini, and S. Vigna. Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks. In *Proc. WWW '11*, pages 587–596, 2011.

[15] P. Boldi, M. Rosa, and S. Vigna. HyperANF: Approximating the neighborhood function of very large graphs on a budget. In *Proc. WWW '11*, pages 625–634, 2011.

[16] P. Boldi, M. Rosa, and S. Vigna. Robustness of social networks: Comparative results based on distance distributions. In *Proc. SocInfo '11*, pages 8–21, 2011.

[17] I. Bordino, P. Boldi, D. Donato, M. Santini, and S. Vigna. Temporal evolution of the UK Web. In *ICDM Work.*, pages 909–918, 2008.

[18] C. Castillo, D. Donato, L. Becchtti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for Web spam. *SIGIR For.*, 40(2):11–24, 2006.

[19] D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-MAT: A recursive model for graph mining. In *Proc. SDM '04*, pages 442–446, 2004.

[20] D. Chakrabarti, R. Kumar, , and A. Tomkins. Evolutionary clustering. In *Proc. KDD '06*, pages 554–560, 2006.

[21] Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng. On evolutionary spectral clustering. *ACM T. Knowl. Disc. Data*, 3(4), 2009.

[22] A. Cuzzocrea, F. Folino, and C. Pizzuti. DynamicNet: An effective and efficient algorithm for supporting community evolution detection in time-evolving information networks. In *Proc. IDEAS '13*, pages 148–153, 2013.

[23] A. Downey. The structural causes of file size distributions. In *Proc. SIGMETRICS '01*, pages 328–329, 2001.

[24] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, pages 17–60, 1960.

[25] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Proc. ACM SIGCOMM '99*, pages 251–262, 1999.

[26] E. Gilbert. Random graphs. *Ann. Math. Stat.*, 30(4):1141–1144, 1959.

[27] M. S. Granovetter. The strength of weak ties. *Am. J. Sociol.*, 78(6):1360–1380, May 1973.

[28] M. S. Granovetter. Economic action and social structure: The problem of embeddedness. *Am. J. Sociol.*, 91(3):481–510, Nov. 1985.

[29] D. Greene, D. Doyle, and P. Cunningham. Tracking the evolution of communities in ′ dynamic social networks. In *Proc. ASONAM '10*, pages 176–183, 2010.

[30] C. Groër, B. Sullivan, and S. Poole. A mathematical analysis of the R-MAT random graph generator. *Networks*, 58(3):159–170, Oct. 2011.

[31] M. Gupta, C. Aggarwal, J. Han, and Y. Sun. Evolutionary clustering and analysis of bibliographic networks. In *Proc. ASONAM '11*, pages 63–70, 2011.

[32] R. A. Hanneman and M. Riddle. *Introduction to Social Network Methods*. U. Cali., Riverside, 2005. http://faculty.ucr.edu/ hanneman/, ret. Mar. 2016.

[33] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. WebBase: A repository of Web pages. *Comp. Networks*, 33(1–6):277–293, 2000.

[34] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *P. Nat'l Acad. Sci.*, 101(Suppl. 1):5249–5253, 2004.

[35] C. Jin, Q. Chen, and S. Jamin. Inet: Internet topology generator. Technical Report CSE-TR-433-00, U. Michigan, 2000.

[36] M. Kim and J. Leskovec. Multiplicative attribute graph model of real-world networks. *I'net Math.*, 8(1–2):113–160, 2012.

[37] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models, and methods. In *Proc. Int'l Conf. Comb. Comput. '99*, pages 1–18, 1999.

[38] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the Web graph. In *Proc. FOCS '00*, volume 57, 2000.

[39] Laboratory for Web Algorithmics. Website. http://law.di.unimi.it/, ret. Dec. 2015.

[40] J. Leskovec. *Dynamics of Large Networks*. PhD thesis, Carnegie Mellon U., Sept. 2008.

[41] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using Kronecker multiplication. In *Proc. ICML '07*, pages 497–504, 2007.

[42] J. Leskovec and R. Sosič. SNAP: A general purpose network analysis and graph mining library in C++. http://snap.stanford.edu/snap, June 2014.

[43] J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graph generation and evolution, using Kronecker multiplication. In *Proc. PKDD '05*, pages 133–145, 2005.

[44] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters, and possible explanations. In *Proc. KDD '05*, pages 177–187, 2005.

[45] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *I'net Math.*, 6(1):29–123, 2009.

[46] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *J. Mach. Learn. Res.*, 11, Mar. 2010.

[47] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Analyzing communities and their evolutions in dynamic social networks. *ACM T. Knowl. Disc. Data*, 3(2), Apr. 2009.

[48] Y. Liu, W.-B. Gong, V. Misra, and D. Towsley. On the tails of web filesize distributions. In *Proc. 39th Allerton Conf. Comm. Control Comput.*, 2001.

[49] J. McAuley and J. Leskovec. Discovering social circles in ego networks. *ACM T. Knowl. Disc. Data*, 8(1), Feb. 2014.

[50] A. Medina, I. Matta, and J. Byers. On the origin of power laws in Internet topologies. *ACM SIGCOMM Comput. Comm. Rev.*, 2(30):18–28, 2000.

[51] S. Milgram. The small-world problem. *Psychol. Today*, 1(1):61–67, 1967.

[52] M. Mitzenmacher. Dynamic models for file sizes and double Pareto distributions. *I'net Math.*, 1(3), 2003.

[53] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *I'net Math.*, 1(2):226–251, 2004.

[54] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cam. U. Press, 2005.

[55] M. Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2):167–256, 2003.

[56] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford U. Press, 1999.

[57] H. Ning, W. Xu, Y. Chi, Y. Gong, and T. S. Huang. Incremental spectral clustering with application to monitoring of evolving blog communities. In *Proc. SDM '07*, pages 261–272, 2007.

[58] G. Palla, A. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[59] V. Pareto. *Cours d'Economic Politique*. Droz, 1896.

[60] D. Pennock, G. Flake, S. Lawrence, E. Glover, and C. Giles. Winners don't take all: Characterizing the competition for links on the web. *P. Nat'l Acad. Sci.*, 99(8):5207–5211, 2002.

[61] R. Rheingans-Yoo. Artificial generation of power-law graphs: A historical survey. Graded assignment, Harvard SEAS, May 2015. http://static.rossry.net/pl_survey.pdf, ret. Dec. 2015.

[62] R. Rheingans-Yoo. Technology for web algorithmics: A historical survey of the interplay between observation and theory. Graded assignment, Harvard SEAS, Dec. 2015. http://static.rossry.net/law_survey.pdf, ret. Dec. 2015.

[63] H. Sedghi, M. Janzamin, and A. Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Proc. AISTATS '16 (to appear)*, 2016.

[64] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, Dec. 1969.

[65] S. Vigna. Stanford matrix considered harmful. *arXiv*: 0710.1962 [cs.IR], Oct. 2007.

[66] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1992.

[67] K. S. Xu, M. Kliger, and A. O. Hero. Evolutionary spectral clustering with adaptive forgetting factor. In *Proc. ICASSP '10*, pages 2174–2177, 2010.

[68] J. Yang and J. Leskovec. Community-affiliation graph model for overlapping network community detection. In *Proc. ICDM '12*, pages 1170–1175, 2012.

[69] J. Yang and J. Leskovec. Overlapping community detection at scale: A non-negative matrix factorization approach. In *Proc. WSDM '13*, Feb. 2013.

[70] J. Yang and J. Leskovec. Overlapping communities explain core–periphery organization of networks. Technical report, Stanford U., Oct. 2014.

[71] J. Yang and J. Leskovec. Overlapping communities explain core–periphery organization of networks. *P. IEEE*, 102(12):1892–1902, Dec. 2014.

[72] G. Zipf. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard U. Press, 1932.

[73] G. Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin Company, 1935.

[74] G. Zipf. *Human Behavior and the Principle of Least Effort*. Addison–Wesley, 1949.

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice, and issued by Adobe in 2007. A template, which can be used to format a PhD thesis with this look and feel, has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/ or from the author at suchow@post.harvard.edu. Graph visualizations were produced with the aid of the D3 library, which can be found online at d3js.org.