Physica A 389 (2010) 4201–4208

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Distance measures for dynamic citation networks

Michael J. Bommarito II [a,b,c,*], Daniel Martin Katz [d,a,c], Jonathan L. Zelner [e,c], James H. Fowler [f,g]

[a] Department of Political Science, University of Michigan, Ann Arbor, United States
[b] Department of Mathematics, University of Michigan, Ann Arbor, United States
[c] Center for the Study of Complex Systems, University of Michigan, Ann Arbor, United States
[d] University of Michigan Law School, United States
[e] Department of Sociology, University of Michigan, Ann Arbor, United States
[f] Department of Political Science, University of California, San Diego, United States
[g] Center for Wireless and Population Health Systems, University of California, San Diego, United States

ARTICLE INFO

Article history:
Received 30 November 2009
Received in revised form 22 April 2010
Available online 11 June 2010

Keywords:
Citation network
Distance measure
Acyclic digraph
Community detection
Clustering
Judicial citations
Dimensionality

ABSTRACT

Acyclic digraphs arise in many natural and artificial processes. Among the broader set, dynamic citation networks represent an important type of acyclic digraph. For example, the study of such networks includes the spread of ideas through academic citations, the spread of innovation through patent citations, and the development of precedent in common law systems. The specific dynamics that produce such acyclic digraphs not only differentiate them from other classes of graphs, but also provide guidance for the development of meaningful distance measures. In this article, we develop and apply our sink distance measure together with the single-linkage hierarchical clustering algorithm to both a two-dimensional directed preferential attachment model as well as empirical data drawn from the first quarter-century of decisions of the United States Supreme Court. Despite applying the simplest combination of distance measure and clustering algorithm, analysis reveals that more accurate and more interpretable clusterings are produced by this scheme.

1. Introduction and motivation

While a variety of algorithms exist for the analysis of undirected or cyclic graphs, e.g., social networks, comparatively little work has been done on acyclic digraphs. The previous literature has focused particularly on the development of canonical random graph models or the application of algorithms for general graphs to this special class [1–4]. While these initial papers have made important theoretical and empirical contributions, investigation into clustering methods for these graphs remains limited.

Dynamic acyclic digraphs arise naturally in the context of document citation networks. In these networks, vertices represent documents and arcs represent the citations from one document to another. Much of the previous literature on citation networks, however, disregards the direction of these arcs (for a notable exception see Leicht et al. [5]). This choice results in undirected graphs with many cycles, and thus allows the application of a wide variety of well-developed algorithms. On the other hand, disregarding direction does discard information about time and the flow of dependency.

Recent work demonstrates that applying methods for undirected cyclic graphs to citation networks may create difficulties [4]. While it is important to identify deficiencies in existing methods, it is more helpful to develop alternative approaches

* Corresponding author at: Department of Political Science, University of Michigan, Ann Arbor, United States. Tel.: +1 5176485688.
E-mail addresses: mjbommar@umich.edu, michael.bommarito@gmail.com (M.J. Bommarito II), dmartink@umich.edu (D.M. Katz), jzelner@umich.edu (J.L. Zelner), jhfowler@ucsd.edu (J.H. Fowler).

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.06.003

designed to properly address these shortfalls. With respect to the context in question, we seek to develop domain-specific methods and measures for citation networks that take the acyclic and directed nature of these networks into account. In this article, we present a novel distance measure that provides better computational efficiency and qualitative accuracy for acyclic digraphs.

2. Properties of dynamic citation networks

2.1. Topological ordering

Since citation networks are dynamic acyclic digraphs, they have a number of important properties that distinguish them from other networks. The most fundamental property of acyclic digraphs is that there exists at least one topological ordering of the vertices [6]. Such a topological ordering can also be used to index the dynamic network G, where G is a nested set of increasing graphs $\{G_1, G_2, \ldots, G_{|V|}\}$. Each G_n is a copy of G_{n-1} with the addition of the nth document of the topological ordering and its corresponding arcs. From this growth dynamic, it is clear that the most natural topological ordering is actually the chronological ordering of the documents.

This ordering implies the existence of another distinguishing property for this class of graphs. Unlike many other growing networks, the set of arcs having non-zero probability at each time step can be explicitly constrained. From a generative framework, this representation acknowledges that arcs cannot assert relationships with unobserved vertices at later times.[1] Formally, such a process evolves on a filtration and can sample from the set of possible arcs at time t given by $\Omega_t^A = \{(x, y) : x \notin V(G_{t-1}), y \in V(G_{t-1})\}$, where $V(G_t)$ is the set of vertices in the graph G_t and t is the index corresponding to the topological ordering. From a statistical framework, in which only the resulting graph is observed, the previous statement can be written $T(x) \leq T(y) \Leftrightarrow \mathbb{P}((x, y) \in A(G)) = 0$, where $T(x)$ is the time that vertex x was introduced into the graph G. This asserts that certain events should not even be considered as possible in statistical models.

2.2. Sinks and dimensionality

A fact that follows immediately from the existence of a topological ordering is that there is at least one document that makes no citations, and at least one document that has never been cited. Documents that contain no citations correspond to nodes with out-degree zero and are called "sinks". The first vertex in the topological ordering is always a sink. Sinks represent documents with no observed dependencies. Thus, with respect to the observed data, they mark the introduction of at least one original or novel idea. Vertices that are not sinks rely on one or more of the ideas provided in one or more sinks.

Though the above conception of citation networks is simple and reasonable, it contradicts patterns often observed in empirical citation data. Namely, many documents contribute novel ideas, but very few feature zero outbound citations. In order to confront this complication and refine our initial conception, it is important to remember documents and citations exist in a high-dimensional space. Documents may contribute novel ideas in one dimension but draw support or comparison from other dimensions—we call these documents "weak" sinks, as opposed to "strong" sinks which make no citations in any dimension.

For a simple but concrete depiction of this problem, consider Fig. 1, a hypothetical subgraph containing vertices a, b, and c, respectively. Vertex a is a "strong" sink, as it features no outbound citations. Vertex b is a "weak" sink, as it cites a on the red dimension but generates no citations with respect to its blue dimension. Vertex c is not a sink, as it relies on b and does not contain the red dimension.

Lacking appropriately granular data, it is often difficult for researchers to separate the dimensions contained within the observable outputs of a given system. However, the above example highlights the specific usefulness of dimensional data. It is important to note that dimensional data is only necessary to identify "weak" sinks but not "strong" sinks. For example, if dimensional data were removed from Fig. 1, thereby removing the coloring of the subgraph, only vertex a would be identified as a sink.[2]

In the context of acyclic digraphs, consider a citation network comprised of linkages between academic articles. While citations to a given article often converge upon a particular dimension or aspect of the work, a given article could be cited on the basis of any of its n dimensions. Building from the example offered in Fig. 1, assume vertex b is an article containing both a novel method and an interesting empirical result.[3] If the *blue dimension* represents the article's substantive topic while the

[1] Draft circulation and pre-print repositories such as arXiv or SSRN allow for the existence of multiple versions of a given document. Given delays between draft publication and subsequent publication, it is possible for triangles or higher-order cycles to exist. It is important for a researcher to consider how best to represent documents that are not consistent with the filtration.

[2] The added returns to differentiating strong and weak sinks will likely vary between applied contexts. However, dimensional data for arcs or vertices would likely improve the quality of the resulting analysis.

[3] While a given article may contain multiple methods or multiple results, for simplicity, assume an article containing a single methodological contribution and a single substantive contribution.

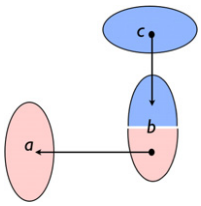


Fig. 1. An example of multidimensional vertex attachment. *a* and *b* are examples of "strong" and "weak" sinks respectively.

red dimension represents the article's methodological contribution, then the basis upon which vertex *c* cites vertex *b* and *b* cites vertex *a* can only be definitively revealed with dimensional data.

While this example is trivial, it reveals a broader property of acyclic graphs. While an author can cite any existing vertex, authors have little control over the basis upon which their work is subsequently cited.[4] One positive feature of the sink method is that it preserves the choices made by the author at the time of authorship.

3. Distance measures

Distance measures between vertices are the predicate to a wide variety of algorithms in network analysis and machine learning [8]. The usage of the terms "distance" and "similarity" are interchangeable in this context. Applications based on the measures presented in this paper can be used in applications requiring similarity measures as well.

As noted earlier, we believe that distance measures employed should incorporate the properties of dynamic citation networks that differentiate them from other classes of graphs. We consider the distance between vertices in the "citation" space, where all documents must orient themselves relative to one or more sinks of information. An appropriate distance measure should decrease as two vertices share more information. The simplest such measure should consider the number of shared sinks between two vertices. Given a vertex i and its set of ancestors A_i, the sinks of i are given by the set $S_i = \{x : \delta^+(x) = 0, x \in A_i\} = S \cap A_i$. Here, $\delta^+(x)$ is the notation for the out-degree of vertex x and S is the set of all sinks of the graph G.

Using this notation, we can represent the distance between vertices i and j as the proportion of sinks they do not share:

$$D_{i,j} = 1 - \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \tag{1}$$

where $|x|$ is the cardinality of set x. Though this distance measure is linearly decreasing in the proportion shared, one can formulate a distance measure from any appropriately decreasing function. The remainder of our distance measures will feature this linear form, but the reader should keep in mind that this is only exemplary.

Furthermore, this measure can be calculated quickly for all pairs of vertices, as its implementation involves little more than graph traversal and set operations. The sets S can be calculated by performing either BFS or DFS from each sink k in the graph and storing k in the set S_i for each i visited in this initial search. The complexity of this step is linear in the number of vertices on a sparse citation network. Next, a naive calculation of $D(i, j)$ for all pairs would require $\binom{|V|}{2}$ set comparisons. By storing component information during the initial search algorithm, however, the number of pairs to be checked can be significantly reduced for most graphs. More sophisticated implementations can achieve even better time complexity by setting some values of $D(i, j)$ within the initial search algorithm. The most naive implementation is quadratic in the number of vertices in the worst case.

In the above distance measure, all sinks are weighted equally. An alternative measure might weight the importance of each sink by the number of unique ancestors shared between vertices i and j that are descended from a sink s of interest. This set of s-ancestors of vertex i is given by $A_{i,s} = \{x : s \in S_x, x \in A_i\} = A_i \cap D_s$, where D_s is the set of descendants of s. This can be interpreted as the ancestors of i who carry the information from sinks s. If even more detail is desired, we might modify the above measure to also incorporate the set of paths from vertices i and j to the sinks of interest. We let $P_{s,i}$ be the set of path tuples $(x_1, \ldots, x_n)$ from s to i. The resulting general equation takes the form

$$D_{i,j} = 1 - \frac{\sum_{s \in S_i \cap S_j} f(A_{i,s}, P_{i,s}, A_{j,s}, P_{j,s})}{\sum_{s \in S_i \cup S_j} f(A_{i,s}, P_{i,s}, A_{j,s}, P_{j,s})}. \tag{2}$$

Straightforward choices of f involve the cardinality of these sets, but some care must again be taken if one desires a distance metric that obeys all axioms. If needed, these functions may take on much more complexity. For instance,

[4] Self-citation is one exception to this rule. See [7].

the importance of a sink might decay as its shortest path length increases. Such fine-grained choices, however, require theoretical justification based on the problem at hand. One should also note that path-based algorithms are likely to exhibit worse time complexity than either of the first two measures.

The above distance measures all bear some similarity to the Jaccard similarity measure, as they involve intersections in the numerator and unions in the denominator [9,10]. However, the Jaccard similarity index only takes into account the neighbors of each node and ignores nodes at any further distance. In comparison, the distance measures presented above may better capture and weight "shared ancestry" than the Jaccard measure.

4. Applications

Once a distance measure has been selected, a number of interesting research questions become relevant. One question of particular interest is whether a given graph exhibits detectable clustering or community structure. A significant amount of recent scholarship has been devoted to designing community detection algorithms for general graphs [11]. In the context of citation networks, there are a number of issues that may impact both the accuracy and longitudinal stability of results produced by traditional community detection methods [4].

One important issue is dimension frequency—that is, some topics may occur much more frequently than others in the overall network. For example, suppose that a vertex z primarily concerns dimension d_1, but also touches upon dimension d_2. If subsequent documents more often confront dimension d_2 than d_1, it is possible that z could receive more d_2-related citations than d_1 citations. As a result, traditional community detection methods are more likely to cluster document z with d_2-related documents than with d_1 documents. Though this example illustrates the way the role of documents within such citation networks may evolve, it is clear that traditional community detection algorithms may produce clusterings that differ from a researcher's specific goals.

For instance, one might seek to cluster documents in a manner consistent with the citation choices of the author at the time the document was written. In this case, sink-based distance measures as presented above might be a good choice for clustering. Take the document z in the example above. Suppose z has three sinks linked to dimension d_1, but only one sink dealing with dimension d_2. Even if many more d_2 documents cite z, they can only share one of four sinks at most. By contrast, d_1 documents can share up to three sinks with z. Thus, regardless of the number of citations from d_1 and d_2 documents, z can still be closer to d_1 documents. Though many gradations of this example exist, when confronted with unnormalized and high-dimensional citation information, sink-based distance measures are likely to be more robust to this issue than traditional community detection methods.

To test whether meaningful clustering can be derived from these sink distance measures, we apply Eq. (1) to two networks below, a theoretical model generated by two-dimensional directed preferential attachment and the other from substantive data offered in the citations of the early United States Supreme Court.

4.1. Comparison on a random model

In Section 2.2, we argue that a number of issues can cause problems with existing community and clustering algorithms. To test this claim, we have generated realizations from a citation model based on two-dimensional directed asymmetric preferential attachment [12].

The model has two types of vertices — red and blue. At each model step, a new vertex is introduced into the network. With probability l_r, a vertex will be red, and thus the complement l_b is the probability of the vertex being blue. To determine how many citations this vertex will make, we sample a uniform random integer between 1 and m. These citation arcs are assigned according to the directed preferential attachment model, where red vertices have probability p_{rr} of citing red vertices and probability p_{rb} of citing blue vertices. Likewise, blue vertices have probability p_{br} of citing red vertices and probability p_{bb} of citing blue vertices. For initial conditions, there are a n_r initial red vertices and n_b initial blue vertices.

In order to demonstrate the problems described above, we choose the parameters of the model to emphasize our example from Section 2.2. The vertex type rates are given by $l_r = \frac{1}{4}$, $l_b = \frac{3}{4}$, the maximum number of arcs per vertex is given by $m = 3$, the preferential probabilities are given by $p_{rr} = 1$, $p_{br} = \frac{1}{4}$, $p_{bb} = \frac{3}{4}$, and the initial number of vertices of each type are given by $n_r = 2$, $n_b = 1$. This models a system with two dimensions where each vertex may only have one dimension, and one dimension occurs much more frequently than another. Furthermore, though one dimension is perfectly homophilic, the other attaches to both. Fig. 2 shows an example realization of this model, where the large squares denote sinks.

To further justify our method, we compare our sink-based approach with directed edge-betweenness [13]. First, we apply Eq. (1) to calculate a full distance matrix for all pairs of vertices. Using this matrix, we then apply a single-linkage hierarchical clustering algorithm to these distances [14]. The resulting dendrogram and its implied clustering are shown in Fig. 3(a) and Fig. 2, respectively. Next, we apply the directed edge-betweenness algorithm to produce the merge dendrogram in Fig. 3(b). Figs. 2 and 3 are both generated from the same underlying network visualized in Fig. 2. The numbering on both figures corresponds to the vertices, and the letters A through E correspond to the clusters detected by the sink method in Fig. 3(a).

The differences in Fig. 3 are striking, though not entirely unexpected. The sink method in Fig. 3(a) identifies five "communities" of vertices at identical branch location. From top to bottom, these branches correspond to (1) vertices that trace back only to vertex 2, (2) vertices that trace back only to vertex 1, (3) vertices that trace back only to vertex 0, (4) vertices that trace back to all three sinks 0, 1, and 2, and (5) vertices that trace back to both vertices 0 and 1.

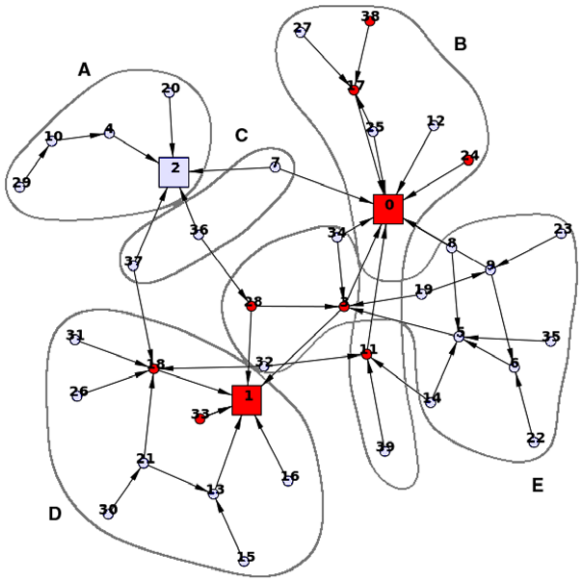


Fig. 2. Realization of a random model with two vertex types and asymmetric attachment probabilities. Clusters implied by the sink method are grouped by gray boundaries.

Since the edge-betweenness algorithm produces binary branching dendrograms like most agglomerative or divisive algorithms, Fig. 3(b) exhibits more complexity than Fig. 3(a). This complexity is sometimes warranted; however, it is often the product of ties in the agglomerative or divisive decision criteria. Since the sink method relies only on hierarchical clustering, it places vertices with equal distance at an equal branch position. In this case, the sink method identifies clusters that are closely related to the underlying network formation process.

4.2. Results for the United States Supreme Court citations

To generate applicability beyond the context of a theoretical model, we applied our approach to the case-to-case citation network of the first quarter-century of decisions of the United States Supreme Court. The structure of this network is of interest to a wide variety of scholars including not only legal academics and social scientists, but also members of the physical science community [15–17,5,18,20]. While it is possible to perform community detection analysis over the total body of Supreme Court decisions, we selected a reduced window of decisions in order to qualitatively examine the results of our algorithm.[5]

The Court's early citation practices indicate an absence of references to its own prior decisions. While the court did invoke well-established legal concepts, those concepts were often originally developed in alternative domains or jurisdictions.[6] At some level, the lack of self-reference and corresponding reliance upon external sources is not terribly surprising. Namely, there often did not exist a set of established Supreme Court precedents for the class of disputes which reached the high court. Thus, it was necessary for the jurisprudence of the United States Supreme Court, seen through the prism of its case-to-case citation network, to transition through a loading phase. During this loading phase, the largest weakly connected component of the graph generally lacked any meaningful clustering. However, this sparsely connected graph would soon give way, and by the early 1820's, the largest connected component displayed detectable structure.

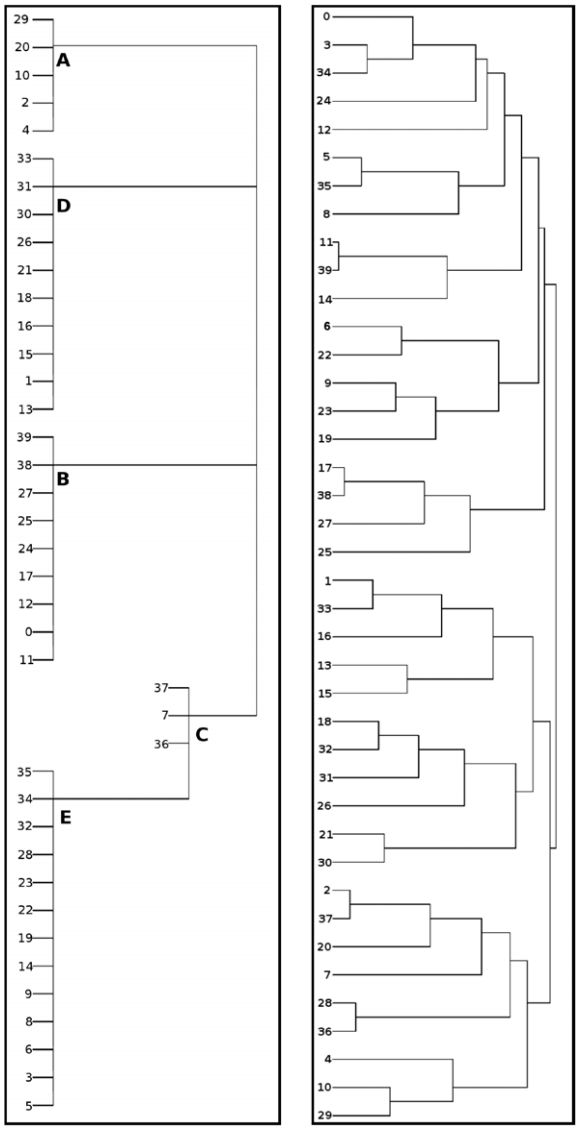
Despite applying naive assumptions about the underlying nature of the data and least complicated clustering algorithm, our qualitative analysis reveals that this scheme produces accurate clusterings. By applying our sink clustering method, we obtain a dendrogram of the network's largest weakly connected component shown in Fig. 4. The coloring in both Figs. 4 and 5 corresponds to two large clusters in the network. Arcs are colored blue or red if the head or tail are of the respective group. Arcs colored green span the two groups.

Documents in both of these colorized clusters engage questions related to maritime and admiralty law.[7] While not a major focus of the docket of the modern court, the early court elaborated a number of important legal concepts through

[5] A number of recent articles have called for more extensive qualitative validation of the clusters and communities detected by such methods [11,19]. In order to address these valid concerns, we have substantively vetted the outputs generated by our method.

[6] The Supreme Court's early jurisprudence references the decisions of England and France as well as several state courts.

[7] It is important to note that some scholars carve a distinction between admiralty (maritime and private international law) and the laws of the sea (public international law). For purposes of characterizing the topical domain, however, we believe it is appropriate to broadly identify these as reasonably related to maritime and admiralty law.



(a) Sink. (b) Betweenness.

Fig. 3. Comparison of dendrograms produced by the sink method and Girvan–Newman edge-betweenness.

the lens of these admiralty decisions. However, despite their general topical relatedness, these two clusters of cases engage substantively different sub-questions, and are thus appropriately divided into separate clusters. For example, the red group of cases engages questions of presidential power and the laws of war, as well as general interpretations of the Prize Acts of 1812. Meanwhile, the blue cluster engages questions surrounding tort liability, jurisdiction, and the burden of proof.

5. Conclusion

We present a novel conception of distance for dynamic citation networks that has trivial implementation and runtime. We successfully apply our sink approach to both a theoretical model and the citation network of the first quarter-century of United Supreme Court decisions. We demonstrate that our method obtains substantively meaningful clusterings and is less susceptible than other clustering methods to common issues resulting from a high-dimensional citation space.

Although the substantive application presented here focuses on the decisions of the US Supreme Court, the applicability of this method is likely not limited to judicial citations. For instance, one could imagine tracing the spread of technological innovation in patent citations or the spread of ideas in a body of academic articles. In future work, we hope to apply this method to such domains, using dimensional data where available. Furthermore, we hope to investigate choices of f in Eq. (2) that match a number of observed phenomena such as the triangle-rich networks seen in Ref. [1].

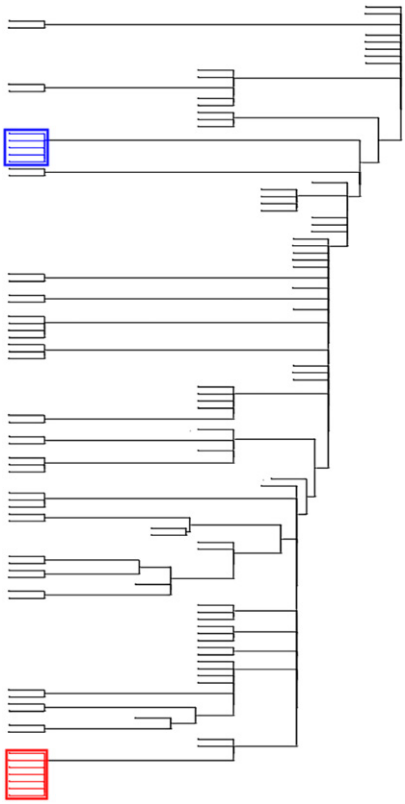


Fig. 4. Dendrogram produced by the sink method applied to the citation network of the first quarter-century of Supreme Court decisions. The outlined groups correspond to the groups indicated in Fig. 5.

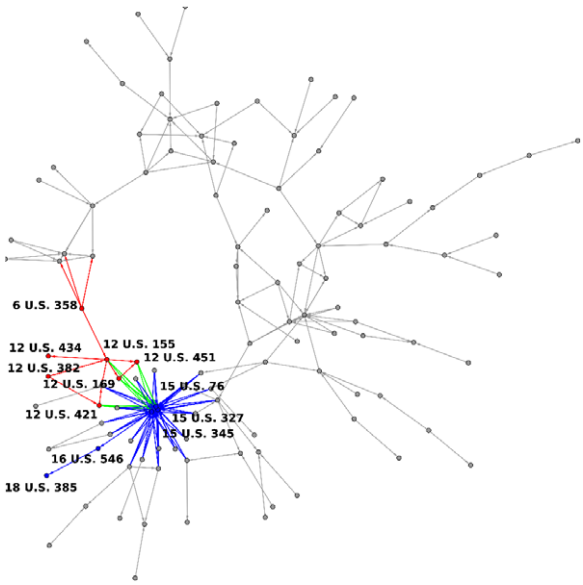


Fig. 5. Largest weakly connected component of the citation network of the first quarter-century of Supreme Court decisions. The vertex and arc colors correspond to the groups indicated in Fig. 4.

Acknowledgements

We would like to thank the Center for the Study of Complex Systems (CSCS) at the University of Michigan and the Michigan Law School for a fruitful research environment. This work was partially supported by an NSF-IGERT fellowship through the Center for the Study of Complex Systems at the University of Michigan, Ann Arbor. We would also like to thank the editor and the anonymous referees for their helpful feedback on the paper.

References

[1] Z.-X. Wu, P. Holme, Modeling scientific-citation patterns and other triangle-rich acyclic networks, Phys. Rev. E 80 (3) (2009) 37101.
[2] B. Karrer, M.E.J. Newman, Random acyclic networks, Phys. Rev. Lett. 102 (12) (2009) 128701.
[3] B. Bollobas, C. Borgs, J.T. Chayes, O. Riordan, Directed scale-free graphs, in: Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2003.
[4] M.J. Bommarito II, D.M. Katz, J. Zelner, On the stability of community detection algorithms on longitudinal citation data, in: Proceedings of the 6th Conference on Applications of Social Network Analysis, 2010.
[5] E.A. Leicht, G. Clarkson, K. Shedden, M.E.J. Newman, Large-scale structure of time evolving citation networks, Eur. Phys. J. B. 59 (1) (2007) 75–83.
[6] J. Bang-Jensen, G. Gutin, Digraphs: Theory, Algorithms and Applications, Springer-Verlag, London, 2000.
[7] J.H. Fowler, D.W. Aksnes, Does self-citation pay? Scientometrics 72 (3) (2007) 427–437.
[8] E. Leicht, P. Holme, M.E.J. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2) (2006) 26120.
[9] P. Jaccard, Etude comparative de la distribution florale dans une portion des Alpes et des Jura, Bull. Soc. Vaud. Sci. Naturel. 37 (1901) 547–579.
[10] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, ISBN: 0-321-32136-7, 2005.
[11] M.A. Porter, J.-P. Onnela, P.J. Mucha, Communities in networks, Notices Amer. Math. Soc. 56 (9) (2009) 1082–1097.
[12] A.-L. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[13] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2004).
[14] F. Murtagh, A survey of recent advances in hierarchical clustering algorithms, Computer J. 26 (1983) 354–359.
[15] T. Smith, The web of the law, San Diego L. Rev. 44 (2007) 309.
[16] J.H. Fowler, T.R. Johnson, J.F. Spriggs II, S. Jeon, P.J. Wahlbeck, Network analysis and the law: measuring the legal importance of precedents at the US supreme court, Political Anal. 15 (3) (2007) 324–346.
[17] J.H. Fowler, S. Jeon, The authority of supreme court precedent, Social Netw. 30 (1) (2008) 16–30.
[18] D. Post, M. Eisen, How long is the coastline of law? Thoughts on the fractal nature of legal systems, J. Legal Studies 29 (2000) 545.
[19] M.E.J. Newman, The physics of networks, Phys. Today (2008).
[20] F.B. Cross, T.A. Smith, A. Tomarchio, The reagan revolution in the network of law, Emory Law J. 57 (5) (2008) 1227.