Community Similarity Metrics in Citation/Influence Network of Philosophers

Justin Payan (s1681453) as part of a group with Alex Hawkins-Hooker (s1675899)

Problem Formulation

In this project, we test community similarity metrics on a network of philosophers. We find two similarity metrics that capture intuition about similarity of the main interests of communities. We also develop a rigorous framework for comparing community similarity metrics, which is not adequately covered in the literature.

A community similarity metric is a function $s: G \times G \rightarrow [0, 1]$ where G is the set of all graphs. Given two graphs X and Y with $|X| \cong |Y|$, s should have the following properties:

1. $s(X, Y) = 1$ when X is isomorphic to Y
2. $s(X, Y) \propto D(L(X), L(Y))$, for some ground truth similarity metric D on the labels $L(X)$ and $L(Y)$.

We also consider functions $f: G \rightarrow [0, 1]$ that can be used to create an s with the above properties via $s(X, Y) = |f(X) - f(Y)|$.

Knowing that two communities are similar along some metric may suggest that they would benefit immensely from communicating with each other. If we find a function f that produces different values for communities and non-communities, it can be used as a community detection algorithm. The community similarity metrics we explore are likely be transferrable to other domains, such as influence networks of historians, authors or entrepreneurs.

Dataset

Our data comes from two sources. We first consider an influence network of philosophers derived from information on DBpedia and Freebase. Undirected edges are drawn from nodes (philosophers) who were sources of influence to nodes who were influenced by that source. We group philosophers that share a “main interest” as being part of the same community. Communities were made connected by simply taking the largest connected component and discarding the other nodes. In previous studies, authors have separated their ground truth communities into multiple communities so that each community is connected [Yang and Lescovec 2012a]. However, because the communities in our network consisted of a large connected component surrounded by isolated nodes and components isomorphic to K_2, we simply deleted the non-connected elements from communities. In addition, we only considered communities with over 7 nodes, since smaller communities would not have enough structure to study. This labeled set of philosophers was called the set of “canonical” philosophers in this work. This graph has 1,530 nodes and 5,546 edges.

We also derive a citation network from the Web of Science, which we call the “contemporary” philosophers. Directed edges are drawn from a citing philosopher to a cited philosopher. This graph has a total of 101,083 nodes and 696,923 edges. A new set of canonical philosophers is identified in this graph as precisely the set of philosophers who were published before 1910. The number of canonical philosophers contained in this graph is 511. These philosophers are in fact a subset of the canonical philosophers from the DBpedia dataset, so we will simply use “canonical” to refer to the DBpedia network for the rest of the paper.

Related Work

There have been some papers that develop exactly the kind of metrics we seek, albeit in different dataset domains. Hadley et al. 2012 discuss a metric called system difference that captures both the level of clustering within a graph as well as the degree distribution and average path length. Koutra et al. 2011 develop a novel graph similarity metric based on belief propagation. Their metric is not applicable to our domain, because they require their networks to have the same set of nodes, with only the edges differing. However, their methodology is instructive, as they compare the scores of 5 different graph similarity metrics and subjectively discuss reasons for the variation in the scores.

A number of papers define metrics on communities in order to rate the effectiveness of community detection algorithms, and such metrics can easily be adapted to comparing pre-defined communities. Yang and Leskovec 2012a discuss 13 different metrics on community structure. They also define four notions of "goodness" of a community, namely separability, density, cohesiveness, and clustering coefficient. Newman and Girvan 2003 also propose metrics on communities. Yang and Lescovec 2012b also present a few metrics on community structure, as well as a general bipartite community affiliation model for graphs. This community affiliation model was the inspiration for our decision to split the dataset into canonical and contemporary philosophers.

Methodology

We compute similarity metrics for every pair of communities in the set of canonical philosophers. We evaluate two functions f: $G \rightarrow [0, 1]$ and three functions s: $G \times G \rightarrow [0, 1]$. One additional random function f: $G \rightarrow [0, 1]$ is used for a baseline comparison. The ground truth for the similarities comes from a combination of the Citation Overlap similarity metric (the most intuitive metric), and a set of rules based on the taxonomy at PhilPapers (http://philpapers.org/browse/all).

Functions f: $G \rightarrow [0, 1]$

- Power law exponent: Fit a power law distribution to the degree distribution of each community individually. The exponent of the power law distribution is the value of f.
- Maximal Internal Community Degree Fraction (ICDF) [Yang and Leskovec 2012b]: The internal degree of a node is defined as the degree of the node in the subgraph induced by the community. Max ICDF is the maximum value of the internal degree of any node divided by community size.
- Random: Computing f consists of simply drawing from a uniform distribution between 0 and 1.

Functions s: $G_1 \times G_2 \rightarrow [0, 1]$

- Citation Overlap: For both graphs G_i, a set c_{ij} is formed for each philosopher (node) j in G_i, as simply the set of all philosophers in the Web of Science citation network who cite node j. Then $C_i = \bigcup_j c_{ij}$. Finally, the metric is computed as $|C_1 \bigcup C_2| / |C_1 \bigcap C_2|$.

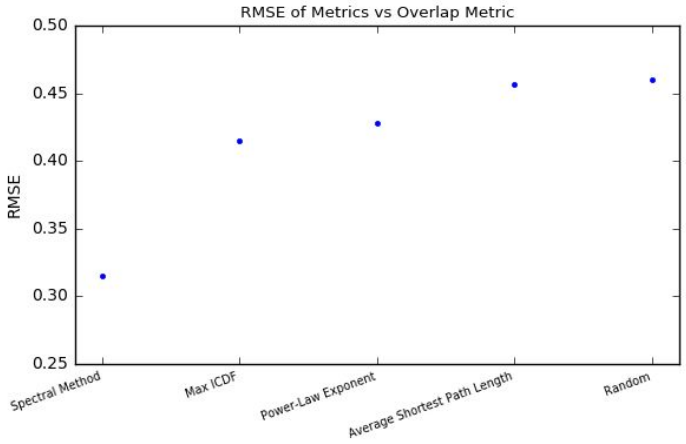
- Average Shortest Path Length: For every node pair (i, j) such that $i \in G_1$ and $j \in G_2$, the distance d(i, j) is computed by calculating the shortest path between i and j. The value of the metric is the average of all such distances.
- Spectral Method: We compute the sorted eigenvalues of the Laplacian for both of the graphs. The similarity is given as the sum of the squared difference of the first k eigenvalues, where k is the smallest number that guarantees 90% of the spectral energy is captured for at least one of the graphs.

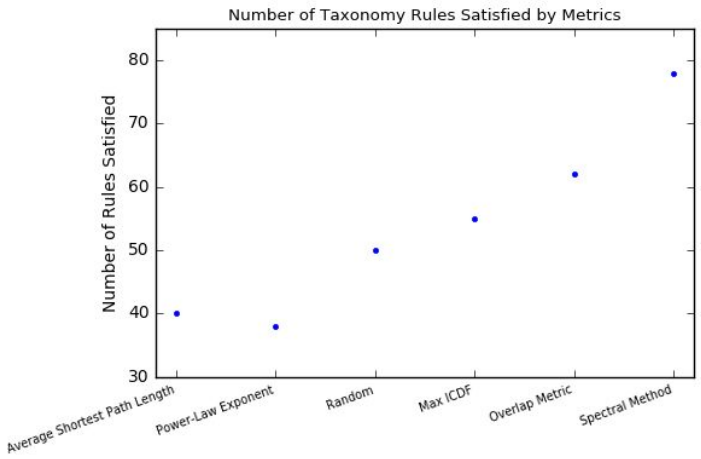
The Average Shortest Path Length metric and the Spectral Method metric must both be normalized after they are computed for all the community pairs.

We also create a system of evaluating metrics that is based on the taxonomy of philosophical topics at PhilPapers. Because the communities were all labelled with topics at the second level of the PhilPapers taxonomy, we were able to group them into three groups based on which root topic they were under at PhilPapers.

The rule-based evaluation essentially just adds a point each time s assigns $s(X, Y) < s(X, Z)$, for X and Y in the same group and X and Z not in the same group (and ordering does not matter for computing s).

Results





Overall, the Spectral Method and Maximal Internal Community Degree Fraction capture intuition about which communities should be similar, while the Average Shortest Path Length and Power Law Exponent are worse than random. We also see that our choice of Overlap Metric as ground truth is well justified.

To evaluate our metrics, we compared their root mean squared error when considering the Overlap Metric as the ground truth. We also computed the number of proper orderings for each of the metrics, as well as for the Overlap Metric.

It is also encouraging that the random similarity metric has the highest RMSE in relation to the ground truth. However, the fact that the average shortest path length and the power law exponent metrics satisfy less taxonomy-based rules than the random metric implies that these metrics actually do not possess sufficient discriminative power on this dataset. Upon reviewing the individual scores for the average shortest path length metric, it is apparent that all scores are between 2 and 3. Thus all communities are about the same distance from each other and the metric can not capture any inherent variation in structure.

The power law exponent similarity metric does vary between about 0.001 and 0.3, which would seem to capture some kind of inherent structural variation. Yet it must be something uncorrelated with the “main interest” of the community.

Although both of the ground truth metrics can be studied in isolation, it is worth noting that the Overlap Metric satisfies the second highest number of taxonomy-based rules. This justifies our use of this metric as a ground truth metric, to an extent. The obvious question is why the spectral similarity method actually satisfies more rules than the overlap metric. This could likely be due to the fact that a handful of canonical philosophers are prolific in a number of different areas which are not at all related. It would be very instructive to remove these prolific canonical philosophers and recalculate the number of rules satisfied by the Overlap Metric.

References

- Hadley, M. W., McGranaghan, M. F., Willey, A., Liew, C. W., & Reynolds, E. R. (2012). A new measure based on degree distribution that links information theory and network graph analysis. *Neural Systems & Circuits*, *2*, 7. http://doi.org/10.1186/2042-1001-2-7
- J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In ICDM, 2012a
- J. Yang and J. Leskovec. Structure and overlaps of communities in networks. In SNAKDD ’12, 2012b

- Koutra, Danai; Ankur Parikh, Aaditya Ramdas & Jing Xiang (2011) Algorithms for graph similarity and subgraph matching. Presented at the Ecological Inference Conference, 17–18 June 2002, Harvard University, Center for Basic Research in the Social Sciences (http://www.cs. cmu.edu/*jingx/docs/DBreport.pdf)
- M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks. Preprint cond-mat/0308217 (2003)