Duhon, Russell J. 2009. Understanding Outside Collaborations of the Chinese Academy of Sciences Using Jensen-Shannon Divergence. Proceedings of SPIE-IS&T Visualization and Data Analysis Conference, San Jose, vol. 7243.

Understanding outside collaborations of the Chinese Academy of Sciences using Jensen-Shannon Divergence

Russell Duhon[1]

Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, 1324 E. 10th Street, Bloomington, IN 47405-3907, USA. rduhon@indiana.edu

Abstract

Using Jensen-Shannon divergence to measure differences in collaboration patterns with outside collaborators makes it possible to understand the structure of those collaborations without direct information about how they collaborate with each other. Applying the approach to data on the outside collaborations of the Chinese Academy of Sciences and visualizing the results reveals interesting structure relevant for science policy decisions.

1. Introduction

Extensive work has been done on understanding the structure of collaborations in co-citation (Bayer, Smart& McLaughlin, 1990; Marshakova, 1973.; Small, 1973; Small& Griffith, 1974; Small& Sweeney, 1985; White& McCain, 1981) and co-author (Newman, 2004; White& Griffith, 1981) networks of papers, authors, and journals. While being continually improved upon (Boyack, Börner& Klavans, 2007; Chen, 1994; Klavans& Boyack, 2006; Small, 2006; Wallace, Gingras& Duhon, Accepted), these approaches require direct evidence to create relationships among the entities. However, many data sets contain interesting information only indirectly evidenced, such as people co-authoring with other people in very similar patterns, but never co-authoring with each other inside the data set. If the primary area of interest is in the relationships directly evidenced by the data, this is not an obstacle, but many times the area of interest is in relationships only implied by the data set. For instance, a lab might be interested in the structure of their collaborations with others, but only have data for their own papers.

A metric projection of Jensen-Shannon divergence provides a distance measure between entities with no clear, direct relationship in the data. The distances reflect the similarity of data items from the perspective of a set of reference data items that they both relate to directly. Unlike raw counts of shared collaborations and measures derived from those, the metric captures all information about how well the collaborator of a paper can be distinguished when it is from one of two collaborators with known distributions. Distinguishing collaborators based on their patterns of collaboration is exactly the type of analysis often desired, making a metric that captures the ability to distinguish such patterns between each pair of collaborators ideal.

This paper discusses reasons for using Jensen-Shannon divergence in clustering and applies the approach to a data set of collaborations by the Chinese Academy of Sciences. The results are shown to reflect known properties of collaboration networks and suggest particular science policy

[1] The author's research is supported by NSF IIS-0535214 award.

considerations. Section two covers Jensen-Shannon divergence, section three discusses related methods for clustering collaboration networks, and section four motivates the approach with applications of other information-theoretic measures to evaluating cluster goodness. From there, section five shows how to apply Jensen-Shannon divergence to data sets like the one analyzed here, and the results of that application are visualized and discussed in section six.

2. Jensen-Shannon divergence

Jensen-Shannon divergence is closely related to Kullback-Leibler divergence, but unlike Kullback-Leibler divergence, which is asymmetric, unbounded, and potentially infinite, it is symmetric and bounded in the interval [0, 1]. Wong and You 1985 (Wong& You, 1985) introduced it, and Lin 1991 (Lin, 1991) discusses many properties. The measure is gaining popularity, and has strong interpretations in multiple fields, from intensive mixture entropy in statistical physics to a log-likelihood ratio in mathematical statistics (Grosse et al., 2002). Most interesting in this case, it can be interpreted as a "capacitory discrimination" measure (Topsoe, 2000). Besides easy interpretation, the square root of Jensen-Shannon divergence is a metric (Endres& Schindelin, 2003; Osterreicher& Vajda, 2003), making it (in metric form) better founded than non-metric measures for use in distance-based algorithms such as agglomerative hierarchical clustering.

The formula for the Jensen-Shannon divergence between two probability distributions p and q is

$$D_{JS}(p,q) = H(.5*p+.5*q) - .5 * H(p) - .5 * H(q)$$

where $H(\cdot)$ is the Shannon entropy of the distribution. It can also be formulated using the Kullback-Leibler divergence, $D_{KL}(p||q)$, as $\frac{1}{2}D_{KL}(p||m)+\frac{1}{2}D_{KL}(q||m)$, where m is $\frac{1}{2}(p+q)$. The result is, in information theoretic terms, the amount of information seeing a symbol generated by a combination of both distributions gives to help figure out the distribution it comes from.

3. Related methods for clustering

Typical approaches to clustering scholarly data include community detection based on graph structure and hierarchical clustering based on distance measures, raw or normalized. Another common approach to discover the structure of a network is pathfinder network scaling, which prunes edges that do not satisfy the triangle inequality from the network, leaving a branching structure only containing the edges not redundant for showing the distances between nodes (Schvaneveldt, 1990).

Community detection that relies on graph structure is good because it avoids transformation from a weighted adjacency matrix to a similarity or distance matrix, which can distort the statistics (Ahlgren, Jarneving& Rousseau, 2003, 2004; Bensman, 2004; Leydesdorff, 2008; Leydesdorff& Vaughan, 2006; White, 2003). However, when there is already a solid distance or similarity metric, very good results can be obtained through procedures like agglomerative hierarchical clustering (Anderberg, 1973; Everitt, 1974). When there is no strong graph interpretation of the data, the case for an approach like agglomerative hierarchical clustering becomes even stronger.

Another direction seeing increasing attention is the use of information-theoretic measures to cluster graphs. For instance, Leydesdorff 2005 (Leydesdorff, 2005) advocates a new approach to graph clustering based on maximizing the in between group entropy. Butte & Kohane 2000 use the mutual information between genes to split genes into clusters by threshholding on that mutual information measure (Butte& Kohane, 2000). Leydesdorff's approach is extremely interesting and powerful, but very new, relatively untested, and computationally expensive. Butte & Kohane's approach is straightforward, but they do not take it very far, possibly because their measure is not a metric.

4. The use of information theoretic measures to evaluate clustering

Information-theoretic measures are also increasingly used to evaluate clusterings. Boyack et al 2005 (Boyack, Klavans& Börner, 2005) follow the comparison technique of Gibbons and Roth (Gibbons& Roth, 2002) in using mutual information to test clusters resulting from several clustering techniques. They use k-means clustering with varying numbers of clusters over layouts done based on several similarity measures to test which similarity measure with their layout algorithm gives the best visual clustering. Gibbons and Roth used the same metric to compare clusters, but in addition to comparing various transformations, they were interested in comparing clustering techniques directly on the transformed data along with Self-Organizing Maps.

Outside of graphs, Grosse et al (Grosse et al., 2002) use Jensen Shannon divergence to evaluate various segmentations of DNA sequences. They find that Jensen-Shannon divergence not only allows stationary and non-stationary sequences to be detected, as they hoped, but also allows coding and non-coding sequences to be detected more accurately than existing techniques. These strong results are significant evidence for the utility of Jensen-Shannon divergence in understanding the distinguishability of groupings. Regarding technique, they find that 'natural' weightings by the proportional lengths of sequences are superior to even weightings.

5. Method

To compute the Jensen-Shannon divergence-based metric for a set of collaborators C, the distributions of their collaborations with a set of researchers R are used. Compute pairwise unweighted Jensen-Shannon divergences and take their square roots to arrive at the related metric. For instance, if collaborator One co-authors with researcher A six times and researcher B four times, the distribution of his collaborations is .6 with A and .4 with B. If another collaborator Two has a distribution of .3 with A and .7 with B, the Jensen-Shannon divergence, by the formula given in section two, will be $H(.5 * [.6, .4] + .5 * [.3, .7]) - .5 * H([.6, .4]) - .5 * H([.3, .7])$, or a little over .046. Despite no direct evidence of their relatedness, this reflects their relatedness relative to their patterns of collaboration with researchers in R.

Unlike Grosse et al 2002, the evenly weighted or unweighted form of Jensen-Shannon divergence is used. This is because the primary interest of this exploration is in distinguishability of patterns of collaboration, and examination of dendrograms from clusterings based on weighted Jensen-Shannon divergence shows that due to large differences in distribution sizes, the clustering does not well reflect this.

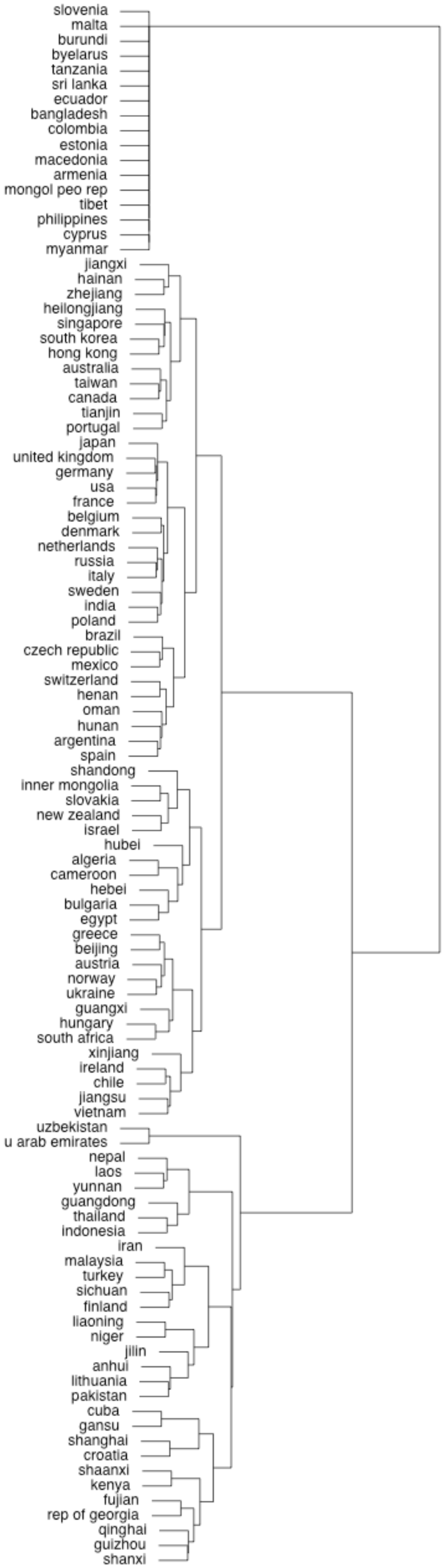
Instead of optimizing on Jensen-Shannon divergence as Leydesdorff does on in between group entropy (which is closely related to Jensen-Shannon divergence), this approach takes advantage of the square root of Jensen-Shannon divergence being a metric. Since it is a metric, techniques well-founded in a metric space, such as agglomerative hierarchical clustering with Ward's method, can be brought to bear. To show the structure of a set of collaboration, perform agglomerative hierarchical clustering using Ward's algorithm. While it is common to cut the resulting dendrogram and arrive at clusters, it will often be instructive to inspect the whole dendrogram, particularly on smaller, easily inspectable data sets.

6. Results and Discussion

The method described above was applied to a set of collaboration data for the Chinese Academy of Sciences (CAS)[2]. The data set contains the number of collaborations by CAS researchers in each provincial level administrative region in China with non-CAS researchers in each provincial level administrative region in China and every country around the world, from 2001 to 2006. For instance, CAS researchers in Shanghai collaborated with researchers in the USA 49 times in that period, and CAS researchers in Yunnan collaborated with researchers in the Netherlands one time in that period.

After creating the Jensen-Shannon divergence based distance matrix between collaborators, agglomerative hierarchical clustering with Ward's algorithm was used to create a dendrogram (Figure 1). The dendrogram has many expected features. For instance, the USA, France, Germany, the United Kingdom, and Japan are all closely clustered. Another expected closely related group contains Nepal, Laos, Thailand and Indonesia. Also, the countries that just collaborate with Beijing, such as Bangladesh and Colombia, are clustered together in a group not closely associated with the other countries.

There are several somewhat more interesting features. For instance, in a grouping that joins one smaller grouping containing Singapore, South Korea, and Hong Kong with another containing Taiwan and Australia, Canada and Portugal are also present. In a grouping closely related to the one containing the USA and Japan, India and Poland appear alongside Belgium, Denmark, the Netherlands, Russia, Italy, and Sweden, but Pakistan and Lithuania are in a different area of the dendrogram entirely from these geographical neighbors.



[2] The data is originally from the Science Citation Index by Thomson Sc
provided by the National Science Library of the Chinese Academy of S

One property of the data set makes certain inferences much easier. The CAS researchers in each provincial level administrative region, as a rule, collaborate much more with non-CAS collaborators in that provincial level administrative region than with non-CAS collaborators in any other provincial level administrative region. This means that countries that cluster near particular provincial level administrative regions also collaborate in a noticeably higher proportion with the CAS researchers in that province. For instance, the grouping including South Korea and Taiwan also includes Jiangxi, Hainan, Zhejiang, Heilongjiang, and Tianjin, provinces in the eastern and southeastern parts of China near, unsurprisingly, South Korea, Taiwan, and many of the others in the same cluster.

Stepping back from the smaller groupings in the dendrogram, the first high level split other than the one splitting off the countries that only collaborate with Beijing puts the collaborators into two groupings. While these groupings are too complex to characterize completely, one grouping definitely contains more developed nations with normalized politics, such as the USA, Austria, Mexico, and Switzerland, while the other includes many less developed nations and ones with less normalized politics, such as Nepal, Thailand, Iran, Turkey, Pakistan, and Cuba. The three top groupings, from slicing the dendrogram at the level that gives three clusters, are shown in Figure 2, which clearly reveals the pattern. Purple regions are in the developed nations grouping, orange regions are in the less developed nations grouping, and green nations are the ones that collaborate only with Beijing.

Figure 1: Dendrogram of CAS Collaborators

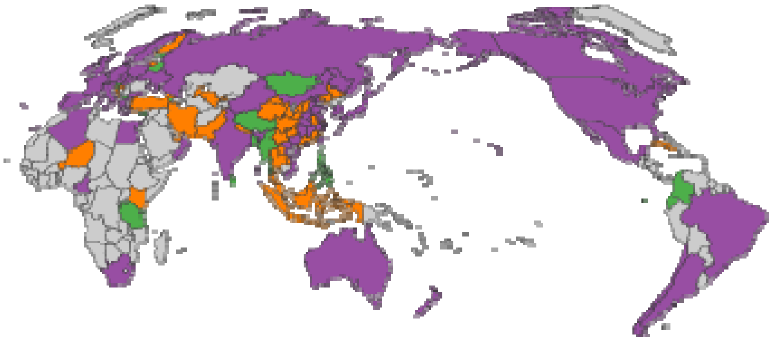


Figure 2: Three Communities

The group of less developed and less normalized nations quickly splits into several smaller groupings, but there are again two fairly interpretable sub-groupings within the developed nations grouping. One contains the larger developed nations for the most part, notably including India as well as the USA, Argentina, et cetera, while the other includes a larger number of smaller nations that also have strong research programs, such as Ireland and South Africa. One of the provincial level administrative regions that appears in this grouping is Beijing. The presence of the provincial level administrative region with the largest number of collaborations and extremely strong ties to Beijing CAS researchers suggests Chinese research collaborations reflect a greater focus on collaboration with these smaller countries compared to their research output. After seeing Beijing, it is natural to ask where Shanghai, the other provincial level administrative region with very large amounts of research, is. Even more surprisingly, Shanghai is in the grouping containing less developed countries!

From a policy perspective, the relationships shown in the dendrogram can be used to characterize the current research patterns of the Chinese Academy of Science, identifying groupings at levels high and low of use for description and discussion. It may also be possible to use the dendrogram to suggest areas worth targeting for new collaborations, particularly multiple collaborations. For instance, given how similarly India and Poland group in the dendrogram, it is worth evaluating possible collaborations involving both nations, especially at a high level, such as a small summit or collective research program, if such collaborations do not already exist.

7. Conclusion

Using a metric projection of Jensen-Shannon divergence to relate collaborators relative to a set of researchers whose activity is well known is a good way to understand the structure of those collaborators. Performing agglomerative hierarchical clustering on the distance matrix yields a dendrogram with strong interpretations from root to leaf that confirm many known patterns and suggest new ones. The relationships uncovered have strong potential use in understanding existing collaborations and suggesting new collaboration opportunities. This technique should be applied to other appropriate data sets, such as the papers originating in individual labs, to better understand how it can be used to uncover new insights.

Other approaches that are well founded given a metric should be explored with this method of relating entities, including common techniques such as multidimensional scaling, pathfinder network scaling, and latent semantic analysis. Additionally, the metric might serve as useful kernel for techniques like support vector machines. The square root of Jensen-Shannon divergence should also be experimented with as a replacement for some common normalizations such as Pearson's R.

8. Acknowledgements

I would like to thank Katy Börner for her help in preparing this paper, and the Chinese Academy of Sciences for providing the data. The author’s work is supported by NSF IIS-0535214 and IIS-0513650 awards.

Bibliography

Ahlgren, Per, Bo Jarneving & Ronald Rousseau. (2003). Requirements For a Cocitation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient. *Journal of the American Society for Information Science and Technology*, *54*(6) 550-560

Ahlgren, Per, Bo Jarneving & Ronald Rousseau. (2004). Rejoinder: In Defense of Formal Methods. *Journal of the American Society for Information Science and Technology*, *55*(10) 936

Anderberg, M. R. (1973). *Cluster Analysis for Applications*. New York: Academic Press, Inc.

Bayer, A. E., J.C. Smart & G.W. McLaughlin. (1990). Mapping Intellectual Structure of a Scientific Subfield Through Author Cocitations. *Journal of the American Society for Information Science*, *41*(6) 444-452

Bensman, Stephen J. (2004). Letter to the Editor: Pearson's r and Author Cocitation Analysis: A Commentary on the Controversy. *Journal of the American Society for Information Science and Technology*, *55*(10) 935-936

Boyack, Kevin W., Katy Börner & Richard Klavans. (2007). *Mapping the Structure and Evolution of Chemistry Research.* Paper presented at the 11th International Conference of the International Society for Scientometrics and Informetrics, Daniel Torres-Salinas & Henk F. Moed (Eds.), Madrid, Spain: Consejo Superior de Investigaciones Cientificas, pp. 112-123.

Boyack, Kevin W., Richard Klavans & Katy Börner. (2005). Mapping the Backbone of Science. *Scientometrics 64*(3) 351-374

Butte, A.J. & I.S. Kohane. (2000). Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements. *Pacific Symposium on Biocomputing*, *5* 415-426

Chen, Chaomei. (1994). *Information Visualization: Beyond the Horizon*. New York: Springer.

Endres, D.M. & J.E. Schindelin. (2003). A New Metric for Probability Distributions. *IEEE Transactions on Information Theory*, *49*(7) 1858-1860

Everitt, B.S. (1974). *Cluster Analysis*. London: Heinemann.

Gibbons, F.D. & F.P. Roth. (2002). Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. *Genome Research*, *12* 1574-1581

Grosse, Ivo, Pedro.Bernaola-Galván, Pedro Carpena, Ramón Román-Roldán, Jose Oliver & H. Eugene Stanley. (2002). Analysis of Symbolic Sequences Using the Jensen-Shannon Divergence. *Physical Review E*, *65*(041905)

Klavans, Richard & Kevin W. Boyack. (2006). Identifying a Better Measure of Relatedness for Mapping Science. *Journal of the American Society for Information Science and Technology* *57*(2) 251-263

Leydesdorff, Loet. (2005). Similarity Measures, Author Cocitation Analysis and Information Theory. *Journal of the American Society for Information Science and Technology*, *56*(7) 769-772

Leydesdorff, Loet. (2008). On the Normalization and Visualization of Author Co-Citation Data: Saton's Cosine Versus the Jaccard Index. *Journal of the American Society for Information Science and Technology*, *59*(1) 77-85

Leydesdorff, Loet & Liwen Vaughan. (2006). Co-occurrence matrices and Their Applications in Information Science: Extending ACA to the Web Environment. *Journal of the American Society for Information Science and Technology*, *57*(12) 1616-1628

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, *37*(1) 145-151

Marshakova, I.V. (1973.). Co-Citation in Scientific Literature: A New Measure of the Relationship Between Publications.". *Scientific and Technical Information Serial of VINITI*, *6* 3-8

Newman, M.E.J. (2004). Who is the Best Connected Scientist? A Study of Scientific Coauthorship Networks. E. Ben-Haim, H. Frauenfelder, Z. Toroczkai, eds. *Complex Networks, Lecture Notes in Physics*, *650* 337-370

Osterreicher, Ferdinand & Igor Vajda. (2003). A New Class of Metric Divergences on Probability Spaces and its Applicability in Statistics. *Annals of the Institute of Statistical Mathematics*, *55*(3) 639-653

Schvaneveldt, R., Ed. (1990). *Pathfinder Associative Networks: Studies in Knowledge Organization*. (Norwood, NJ: Ablex Publishing.

Small, Henry. (1973). Co-Citation in Scientific Literature: A New Measure of the Relationship

Between Publications. *JASIS*, *24* 265-269

Small, Henry. (2006). Tracking and Predicting Growth Areas in Science. *Scientometrics*, *63*(3) 595-610

Small, Henry & Belver C. Griffith. (1974). The Structure of Scientific Literatures I: Identifying and Graphing Specialites. *Science Studies*, *4*(1) 17-40

Small, Henry & E. Sweeney. (1985). Clustering the Science Citation Index® Using Co-Citations: I. A Comparison of Methods. *Scientometrics*, *7*(3-6) 391-409

Topsoe, F. (2000). Some Inequalities for Information Divergence and Related Measures of Discrimination. *IEEE Transactions on Information Theory*, *46*(4) 1602-1609

Wallace, M.L., Yves Gingras & Russell J. Duhon. (Accepted). A New Approach for Detecting Scientific Specialites from Raw Cocitation Networks. *Journal of the American Society for Information Science and Technology,*

White, Howard D. (2003). Author Cocitation Analysis and Pearson's r. *Journ of the American Society for Information Science and Technology*, *54*(13) 1250-1259

White, Howard D. & Belver C. Griffith. (1981). Author Cocitation: A Literature Measure of Intellectual Structure. *Journal of the American Society for Information Science 32* 163-172

White, Howard D. & Katherine W. McCain. (1981). Author Co-Citation: A Literature Measure of Intellectual Structure. . *Journ of the American Society for Information Science*, *32* 163-172

Wong, A & M. You. (1985). Entropy and Distance of Random Graphs with Application to Structural Pattern Recognition. *Pattern Analysis and Machine Intelligence*, *7* 599-609