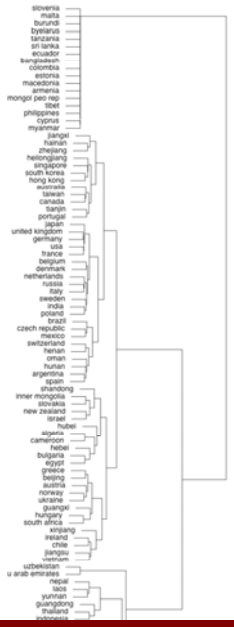
Understanding outside collaborations of the Chinese Academy of Science using Jensen-Shannon divergence



Visualization and Data Analysis 2009
San Jose, California, USA
19 January 2009

Russell J. Duhon
rduhon@indiana.edu
Cyberinfrastructure for Network Science Center
School of Library and Information Science
Indiana University
Bloomington, IN, USA

Table of Contents

Collaboration

What is Collaboration?

Collaboration

What is Collaboration?

Good question . . .

but a lot of collaboration is pretty easy to recognize. Co-authorship is a common sort looked at in scientometrics.

Collaboration

How do we measure collaboration?

In a lot of ways, some of which we've heard about earlier today. Returning to co-authorship, one way is to make networks and analyze those.

Collaboration

What if we don't have a nice co-authorship network?

In other words, how do we uncover less obvious characteristics of interest?

Entropy

Entropy

What if we don't have a nice co-authorship network?

In other words, how do we uncover less obvious characteristics of interest?

Entropy

Many conceptualizations, and in one sense helps measure distinguishability.

Entropy

Shannon Entropy

$$-\sum p(x_i)\log p(x_i)$$

Entropy

Jensen-Shannon Divergence

$$H(.5p + .5q) - .5\,H(p) - .5\,H(q)$$

where *H* is the Shannon Entropy.

Collaboration and Entropy

So, given how a set of collaborators co-author with some control set, we can treat those as distributions and uncover the distinguishability of those distributions.

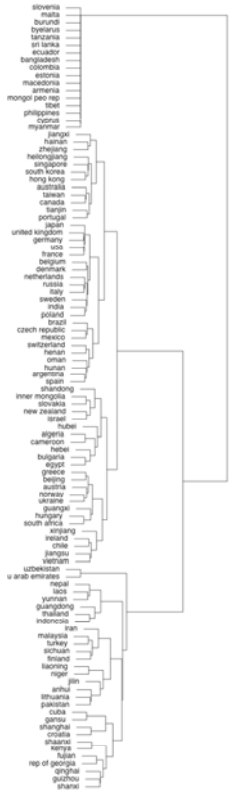
Collaboration and Entropy

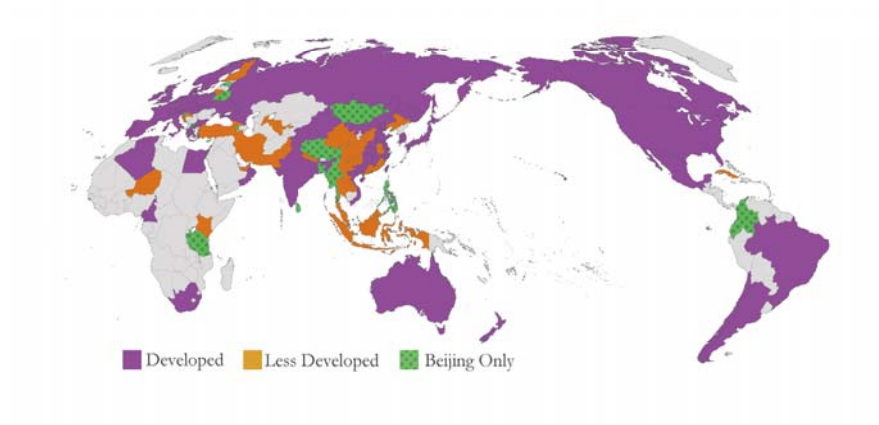
Then, given a pairwise distinguishability metric, algorithms taking advantage of metric similarity become available.

Collaboration and Entropy

Take those pair-wise metrics, and do Agglomerative Hierarchical Clustering with Ward's Algorithm to look for structure.

Information

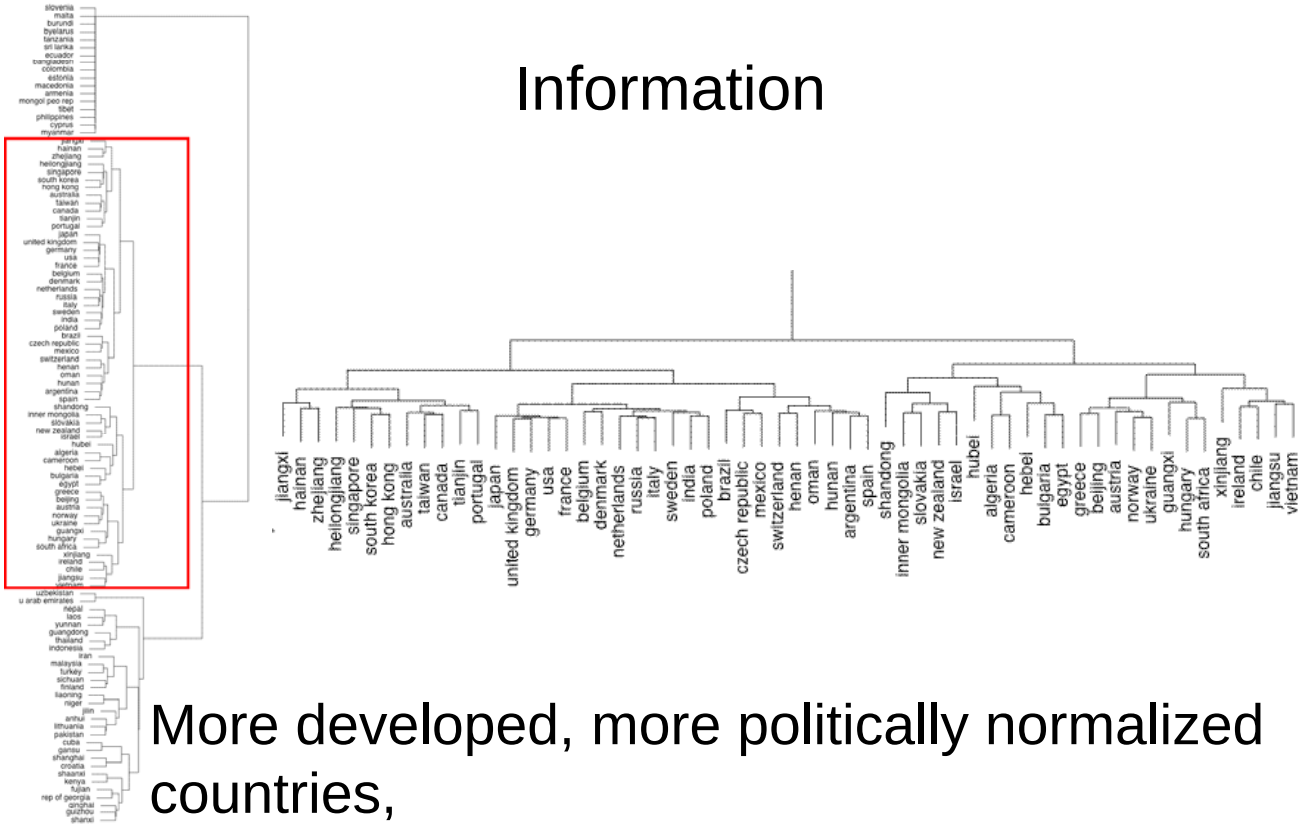




Information

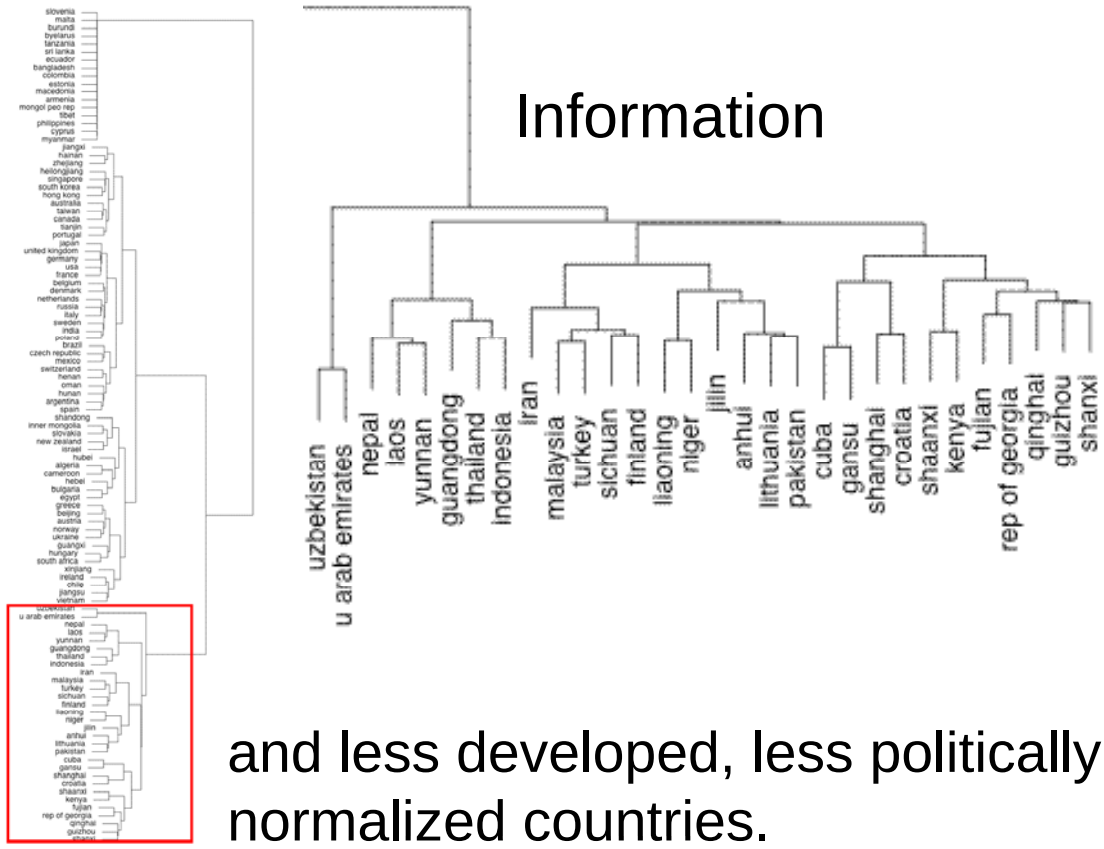
Ignoring Beijing-only collaborators, there are two primary clusters:

Information



More developed, more politically normalized countries,

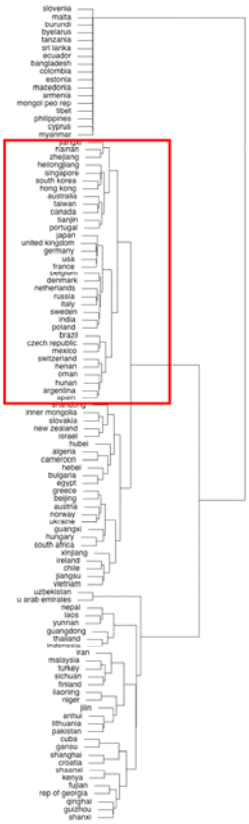
Information



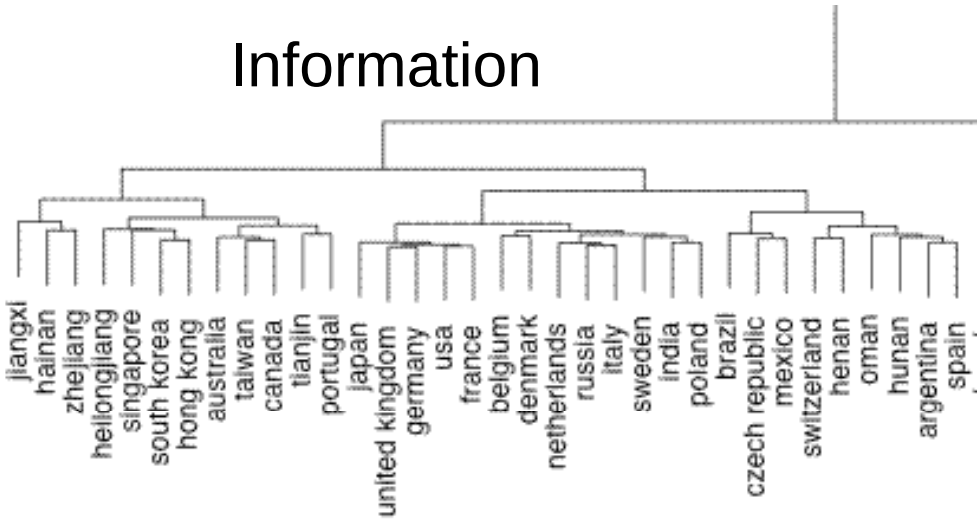
and less developed, less politically normalized countries.

Information

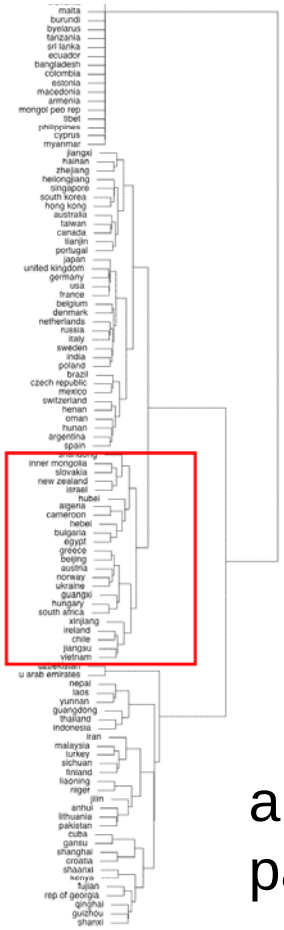
Within the cluster of more developed, more politically normalized countries, there are again two clear clusters:



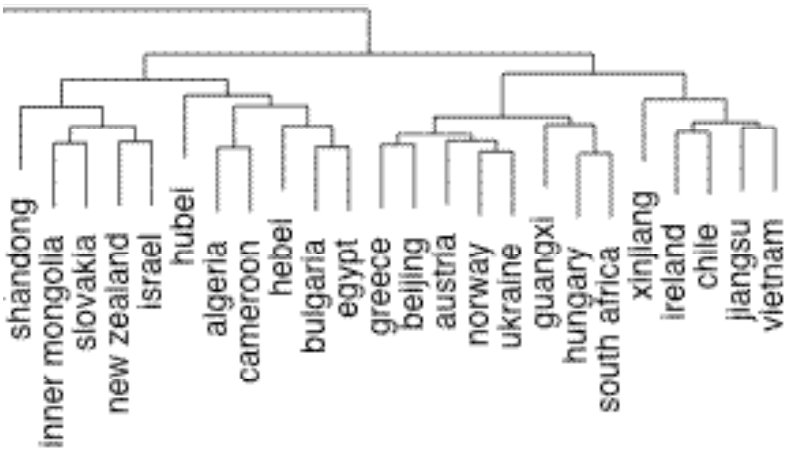
Information



Larger, more developed countries, including the up-and-coming BRIC members,



Information

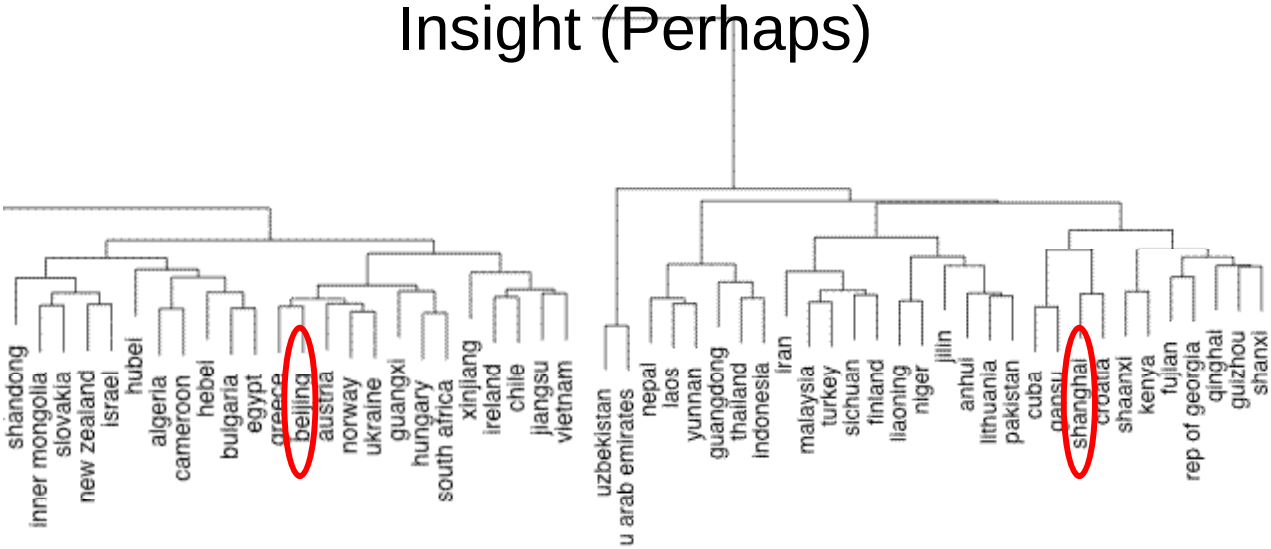


and much smaller countries, for the most part

Information

Looking more closely, there are many small clusters that show expected patterns.

Insight (Perhaps)



Starting to approach insight, the locations of Chinese collaborators hint at potential explanations.

Insight (Perhaps)

One of the strongest potential applications will be identifying possible sites for the application of policy tools to strengthen, sustain, and transform collaborative structures.

Applications in the Small and Large

Remember the motivation: understanding structure when there is less information than ideal.

Applications in the Small and Large

Small research groups generally know who their members collaborate with and how much, but rarely how those people collaborate with each other.

Applications in the Small and Large

Large organizations often have similar information limitations, particularly due to identification and data dirtiness.

Applications in the Small and Large

In both cases, this procedure may be very useful. The insights hint at potential relations among outside collaborators, and comparison of distributions minimizes the impact of dirty data.

Thank You

Russell Duhon
rduhon@indiana.edu

NSF IIS-0534909 and IIS-0513650 awards