Measuring Collaborative Distance

EE126 Fall 2014 | Akshay Narayan, Nishanth Mohan, Japheth Wong

Introduction

PageRank is an algorithm used by Google to rank the importance of pages in its search results. Although the algorithm works well for indexing the web, it is able to provide only a global ranking over the data. Using PageRank alone, it is difficult to investigate local relationships in data and rank by metrics according to a given perspective, or starting node in the graph.

One such relationship that we believed is interesting is the concept of a collaborative distance on published works. One popular example of this is the Erdős number, a number which measures the collaborative distance between an individual with the mathematician Paul Erdős. We believed it would be both interesting and amusing to bring this concept closer to home and compute a "Ramchandran number", a measure of the collaborative distance with Professor Kannan Ramchandran based on the papers available on Google Scholar.

The challenge with this, however, is that researchers tend to collaborate with a high number of individuals. If we model the coauthor relationship with a graph, we would expect each node of the graph to have a high degree. Given the number of researchers covered by Google Scholar, traversing this entire graph to compute the true Ramchandran number is computationally expensive and impractical. Instead, we believed we could make use of the concepts in EE126 to approximate an individual's Ramchandran number.

Algorithm

We start by populating an initial database of authors and an estimate of their Ramchandran numbers. Viewing the "Ramchandran number graph" as tree rooted at Kannan Ramchandran, we use a depth limited breadth first search to compute Ramchandran numbers of various authors closest to him. For a particular author, we sample a random subset of his co-authors (based on an initial branching factor) from Google Scholar, compute his respective Ramchandran number and run this process recursively on each of these co-authors.

Once this initial database has been computed, we pick a random author from the initial database, sample a random subset of his co-authors, and update their respective Ramchandran numbers. However, instead of considering each of the co-authors as before, we pick a random one and run the same update process on him. While this results in relatively long chains of authors who have large Ramchandran numbers, it has the added benefit of covering the space of authors more widely and quickly than the initial breadth-first search.

Challenges

One issue we were worried about is that Google throttles queries: we needed to make sure we could get a representative view of an individual's Ramchandran number without having to run too many queries. In addition, we currently depend on Google Scholar to maintain a consistent representation for each author, i.e. "Kannan Ramchandran" would always be represented as "Kannan Ramchandran".

Results

We used an initial BFS branching factor of 5, initial BFS depth of 4, random traversal branching factor of 10, and random search depth of 150. We have included the top 20 and last 5 results below; a complete list of the results can be found in our file, results.txt.

Name	Ramchandran Number
Kannan Ramchandran	0
Prakash Ishwar	1
Gerald Friedland	1
Longbo Huang	1
Kv Rashmi (Rashmi Vinayak)	1
Minghua Chen	1
Keith Ross	2
Krste Asanovic	2
Brian Kingsbury	2
Osama Khan	2
Ulas Kozat	2
Zeev Dvir	2
Nicholas Ja Harvey	2
Jin Li	2
Steve Renals	2
Parikshit Gopalan	2
Mihai Patrascu	2
Zongpeng Li	2
Nikki Mirghafori	2
Chuck Wooters	2
Ness Shroff	2
…	
Preslav Nakov	17
Andrew J. Cowell	18
Ted Briscoe	18
Keith Fligg	19

Future Work

We leave for future work the ability to find the Ramchandran number of an arbitrary researcher, rather than just searching for and assigning Ramchandran numbers to individuals in the co-author graph. This could be accomplished by first searching using the algorithm described above for the researchers with the lowest Ramchandran numbers. Once this dataset is established, finding the Ramchandran number of an arbitrary individual P would involve a loop-free random graph traversal starting from P until someone with a known Ramchandran number X was found. By running this random search multiple times, a reasonably accurate estimate for the additional collaborative distance between X and P could be found. Then, an approximation for the Ramchandran number of P would be the Ramchandran number of X plus the additional collaborative distance.