Thesis: An Analysis of the Friendship Paradox and Derived Sampling Methods

Abstract: The friendship paradox is the phenomenon that people's friends, on average, have more friends than they themselves do. This phenomenon was the inspiration for a well-known sampling method where a randomly sampled vertex is exchanged for one of its randomly sampled neighbors in order to increase the expected degree of the selection.

Our research explores the friendship paradox on a local level. We introduce a metric, the friendship index, that quantifies the effect of the friendship paradox on an individual vertex. We demonstrate significant usefulness for this metric in both its original form and as the basis for aggregate measures of a graph. In particular, we show its similarity to the famous degree-homophily measure, assortativity, as well as vertex/graph characteristics that it reflects, but which are not captured by assortativity.

We then explore the random neighbor sampling method. We connect this method to the degree-homophily trait of a graph that is captured by the friendship index and assortativity. We introduce a cost model that explores random neighbor sampling from the perspective of the extra costs, computational or practical, that it incurs, and we reframe its strength in the cost-benefit analysis we perform. Our cost model leads to a number of interesting tweaks that improve the method while targeting specific costs that might need to be minimized in a given situation.

One of these tweaks suggests a very significant modification to random neighbor sampling, inclusive random neighbor sampling, where we retain the first vertex and select it if it is of higher degree than the sampled neighbor. We also apply the idea of inclusivity to another, lesser-known sampling method, random edge. In inclusive random edge sampling, an edge is sampled instead of a vertex, and the endpoint with higher degree is selected. Our research brings out many interesting theoretical results related to the original versions of the sampling methods as well as the new inclusive ones. We highlight numerous characteristics that affect the performance of the methods and conduct various experiments to demonstrate their strengths and weaknesses. Among many significant results, we find that very often inclusive random edge is the strongest method, despite the fact that random neighbor is the more popular of the exclusive versions. This suggests that tracking edges may be a worthwhile addition to networks where high-degree random sampling is required.

Committee:

- Professor Amotz Bar-Noy, Mentor, Brooklyn College
- Professor Ted Brown, Queens College
- Professor Saad Mneimneh, Hunter College