TESTING GRAPH CLUSTERABILITY: A DENSITY BASED STATISTICAL TEST FOR DIRECTED GRAPHS

Houyem Demni[1], Pierre Miasnikof[2], Alexander Y. Shestopaloff [3] , Cristián Bravo [4] and Yuri Lawryshyn [2]

[1] Department of Economics and Law, University of Cassino and Southern Lazio, Cassino, Italy, (e-mail: houyem.demni@unicas.it, mariorosario.guarracino@unicas.it)

[2] University of Toronto, Toronto, ON, Canada, (e-mail: p.miasnikof@utoronto.ca, yuri.lawryshyn@utoronto.ca)

[3] Queen Mary University of London, London, UK and Memorial University of Newfoundland, St. John's, NL, Canada (e-mail: a.shestopaloff@qmul.ac.uk)

[4] University of Western Ontario, London, ON, Canada (e-mail: cbravoro@uwo.ca)

ABSTRACT: In this work, we extend a recent statistical test for graph clusterability to directed graphs. Graph clustering, or network community detection, is a pivotal topic in network science. It consists of labeling nodes so they form subsets that display a greater similarity to each other than to the remaining vertices on the graph. Here, node similarity is measured in connection probability or edge density. Similar nodes have a greater connection probability to each other than to other vertices. However, not all graph have a clustered structure. While the goal of graph clustering is to offer a meaningful summary of a graph through vertex clusters, not all graphs can be summarized in this way. In cases where a graph is not clusterable, clustering is not only a waste of time, it inevitably leads to misleading conclusions. We tailor a statistical test developed for undirected networks to directed ones. The test is based on measuring the heterogeneity of local densities. It does not assume any particular graph generative model or edge probability distribution. The test only rests on the hypothesis that a clusterable graph must display a mean local (induced subgraph) density that is significantly greater than the graph's overall density. We posit that this inequality is a necessary (but not sufficient) condition for a graph to have a clustered structure. After highlighting the probabilistic nature of local and global densities, we offer a statistical test to assess the significance of this inequality in densities. This test is also based on sampling node neighborhoods and is thus well suited to very large data sets. We have validated our test on several synthetic graph structures and real world networks. We have also compared our test to other recent statistical tests. Our findings show that our test is more responsive to networks structure than its alternatives.

KEYWORDS: Clustering, global densities, local densities, networks.