Atelier
« Défis statistiques et computationnels dans les réseaux et la cybersécurité »
04-08 mai 2015

Workshop
"Statistical and computational challenges in networks and cybersecurity"
May 04–08, 2015

Comparison measures and an ensemble method for large-scale graph clustering

François Théberge*

theberge@ieee.org

The problem of large-scale graph clustering is important for a variety of applications including relational data exploration, visualization, community detection, and cyber defence. In the cyber defence context, one objective is to complement signature-based detection schemes with behaviour-based anomaly detection. However, the size of the graphs involved often requires some pre-processing to be done, such as graph clustering, in order to fit within computational and/or storage limits.

Several graph clustering algorithms have been proposed and studied over the years, including our own which is based on the ensemble learning paradigm. Our goal was to develop a method that is both scalable and robust, where the clusters can be used for tasks such as finding strong communities and seed set expansion.

While evaluating graph clustering algorithms for application to our problems, it became clear that the choice of measure used to compare results can have a huge impact on the conclusions. It is common practice to use set partition measures on the vertices of a graph to compare clustering results, thus neglecting the overall graph structure. We review several such measures based on information-theory or pairwise-counting. All measures run into bias when comparing clustering results having different cardinalities, and corrections exist for the information-theory based measures. We observe that those can be extended to most pairwise-counting measures. We also introduce some graph-aware measures as complements to the set-partition ones. Graph clustering can also be viewed as binary edge classification, which allows us to define such graph-aware measures. We describe some properties of those measures, and we illustrate the complementarity of set-partition and graph-aware measures via empirical results.

*Département de mathématiques et de statistique, Université d'Ottawa, 585 King Edward, Ottawa, ON K1N 6N5, CANADA.