On clustering heterogeneous networks

Forough Poursabzi Sangdeh and Ananth Kalyanaraman
School of Electrical Engineering and Computer Science,
Washington State University, Pullman, WA 99164
E-mail:*f.poursabzisangdeh@wsu.edu, ananth@eecs.wsu.edu*

1 Introduction

One of the primary contributing factors for increasing complexity in network analysis is data heterogeneity. Multiple types of data are available from a diverse range of sources and scientists need new capabilities to integrate information from these diverse data types in order to gain a deeper and holistic understanding of a naturally built system. In this paper, we visit the problem of *clustering networks built out of heterogeneous data*. Given a k-partite network, where each of the k vertex partitions represents a distinct vertex (data) type and the edges represent inter-type relationships, the problem is one of identifying subsets of vertices of one type that are strongly linked to subsets of vertices of one or more of the other types. While the idea of modeling heterogeneity as a graph problem is not new, algorithm development efforts have been more recent[1,2,3]. While most of these methods have shown appreciable efficacy in clustering or scaling, they all make one or both of the following assumptions: that the number of output clusters is known *a priori*; or that the network has a particular topology. On the contrary, many real world use-cases exist (e.g., Figure 1a) where the number of output clusters is unknown and the network can be of an arbitrary k-partite structure with or without intra-type edge information.

2 Key ideas and preliminary evaluation

In this paper, we present some of our key ideas to overcome the above challenges.

Formulation: First, to make it broadly applicable to a wide range of real world networks from different scientific domains, we treat the problem to be one of *co-clustering* an *arbitrary* k-partite network, with no restrictions on either the network topology or the value of k. A "*co-cluster*" is defined to be a collection of at most k non-empty subsets of vertices, each from a distinct partition, such that for every subset in the collection there exists at least one other subset to which it is "strongly linked". The notion of strong linkage between two subsets is a measure of the relative intensity of edges flowing between the two subsets compared to the edges exiting the subsets. *Not* enforcing that the subsets of a co-cluster necessarily cover *all* the k partitions provides a provision to accommodate for potential weak linkages or an absence of linkage information. The formulation also has the flexibility to allow for potential overlaps between reported co-clusters to capture several real world cases. To allow for the incorporation of any available intra-type edge information, those vertex partitions that have intra-type edge information can be duplicated. For instance, if $m \leq k$ partitions have intra-type edge information, our formulation conceptually treats the input as an $(m+k)$-partite graph with the intra-type edge connecting the vertices of the corresponding duplicated partition.

Algorithm: Presently, we are investigating an algorithmic approach to identify co-clusters from an arbitrary k-partite network under the formulation presented above. At a high level, the approach taken is to identify clusters within individual partitions (labeled as "targets") as imposed by all

[1]B. Gao *et al.* In Proc. SIGKDD, pp.41-50 (2005).
[2]M. Hartsperger *et al.* BMC Bioinformatics 11, 522 (2010).
[3]Y. Sun *et al.* In Proc. ACM SIGKDD pp. 797-806 (2009).

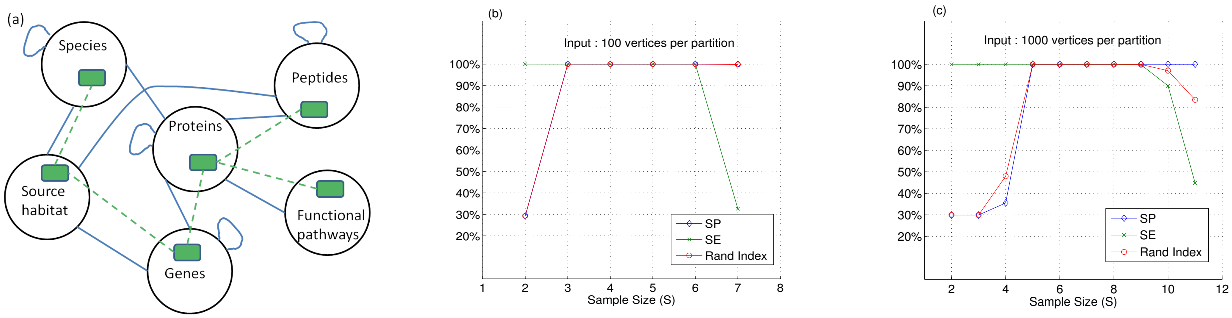

Figure 1: Part (a) shows a systems biological 6-partite network containing both inter- and intra-type edges and also an illustration of a sample co-cluster (shown connected by dotted lines). Parts (b) and (c) show the results of comparing our results against the expected results of a synthetic 4-partite network benchmark with predefined cluster structure. The qualitative measures plotted are as follows: Specificity (SP)= $\frac{TP}{TP+FP}$, Sensitivity (SN) = $\frac{TP}{TP+FN}$, and Rand Index $=\frac{TP+TN}{TP+TN+FP+FN}$, where True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) are measures of agreements and disagreements in vertex pair clustering between the two schemes (test and truth).

other neighboring partitions (labeled as "attributes"), and then combining them into co-clusters so as to maximize an objective function — the details are beyond the scope of this paper. Here, we expand on a key step of the approach which aims to identify an initial set of clusters within each target partition as imposed by all its attribute partitions. To implement this step, we are exploring the use of a randomized heuristic[4] based on the following simple observation: that two vertices of a target partition are likely to belong to the same cluster if they share a significant fraction of their neighborhood within most (if not all) of the other neighboring partitions. In other words, the probability that two vertices are clustered together is proportional to the ratio of the cardinalities of their neighbor-sets' intersection over union. The Minwise Independent Permutations theory[4] provides a quick way to evaluate this ratio by generating fixed size samples from both sets and comparing those samples, and this was previously exploited under unipartite settings[5].

We extend this approach under k-partite settings as follows: Let V_t denote a target partition and V_a denote any of its attribute partitions. To compute a grouping of vertices in V_t as imposed by a given V_a, we implemented the following approach: Identify all pairs of vertices in V_t that share at least one sample in V_a generated over multiple random trials. Create clusters by performing a transitive closure union of vertices based on these pairs (using union-find data structure). Subsequent combination of clusters imposed by different partitions is computed by performing an intersection of the clusters, resulting in the final grouping of vertices within the target partition.

Experimental evaluation: To test the clustering efficacy of the above approach, we constructed a synthetic 4-partite network of different sizes (hundred to thousand vertices per partition) with manually pre-defined co-cluster structures using pre-defined probabilities for the intra-cluster edges. We then applied our test implementation on these input networks and compared the clustering obtained for a designated target partition against its expected clustering. Figure 1b and c summarize qualitative results of our method. As can be observed, the quality of the clustering depends on the sample size (S) — if S is too small, the approach over-merges resulting in high false positive rate, whereas a large value degrades sensitivity. There is a clear range of sample size values ([3,6] for *input*_100 and [5,9] for *input*_1000) for which our clustering was 100% in agreement with the benchmark results, demonstrating the suitability of the approach for use in overall co-clustering process. However, further development and more testing on more complex benchmarks with different degree and cluster distribution characteristics is required to refine its efficacy for larger-scale real world networks.

[4] A.Z. Broder *et al.* In Proc. STOC, pp. 327-336 (1998).
[5] D. Gibson *et al.* In Proc. VLDB, pp. 721-732 (2005).