



Stochastic online clustering of massive graphs

Satu Virtanen[†]

Laboratory for Theoretical Computer Science
Helsinki University of Technology, Finland
email: satu@tcs.hut.fi

Most graph-theoretical clustering algorithms are global, i.e., require the complete adjacency relation of the graph representing the examined data. We propose instead a local approach that computes clusters in graphs, one cluster at a time, relying only on the neighborhoods of the vertices included in the current cluster candidate. The clusters may be identified either by complete search, or when approximate accuracy is sufficient, employing heuristic methods.

Graph clustering

Clustering is the process of organizing data into meaningful groups in order to interpret some properties of the data; we consider data that consists of a set of vertices that are connected by a set of edges, where an intuitive cluster is a subset of vertices sharing relatively many edges with respect to the global structure.

- In a graph $G = (V, E)$, a cluster candidate is a set of vertices $C \subseteq V$.
- The set of edges of the subgraph induced by a cluster is $\mathcal{E} = \{(u, v) \in E \mid u, v \in C\}$.
- The size of the cluster $|C|$ is the number of vertices included in it.
- The internal degree and outbound degree of C are

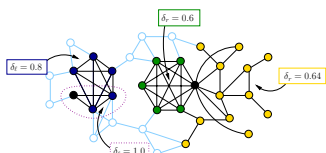
$$\begin{aligned} \text{deg}_{\text{int}}(C) &= |\mathcal{E}| = |\{(u, v) \in E \mid u, v \in C\}|, \\ \text{deg}_{\text{out}}(C) &= |\{(u, v) \in E \mid u \in C, v \in V \setminus C\}|. \end{aligned}$$

- Possible and commonly used cluster fitness measures

$$\text{Local density} \quad \delta_\ell(C) = \frac{|\mathcal{E}|}{\binom{|C|}{2}} = \frac{2|\mathcal{E}|}{|C|(|C|-1)}$$

$$\text{Relative density} \quad \delta_r(C) = \frac{\text{deg}_{\text{int}}(C)}{\text{deg}_{\text{int}}(C) + \text{deg}_{\text{out}}(C)}$$

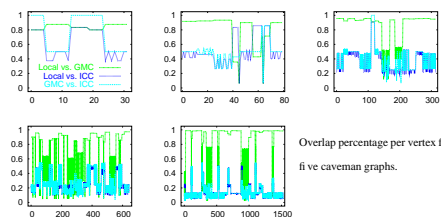
δ_ℓ measures the “dependence” of the cluster members of each other, whereas δ_r measures the “independence” of the vertex set C from the rest of the graph. Good clusters have both high $\delta_\ell(C)$ and high $\delta_r(C)$; optimizing δ_ℓ prefers small cliques over larger but slightly sparser subgraphs, and clusters with optimal δ_r may be sparse.



Examples of undesirable behavior of the measures δ_ℓ and δ_r .

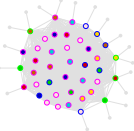
Comparison with other methods

We compared the clusterings achieved with the local method to the clusterings of GMC (Geometric Minimum Spanning Tree Clustering) with additional linear-time post-processing and ICC (Iterative Conductance Cutting) [6] for five caveman graphs of different size. For each graph, we compared the clusters of each vertex achieved with the three methods by calculating how many percent of the vertices of the smaller cluster are also included in the larger one.



Overlap percentage per vertex for five caveman graphs.

The post-processed GMC and the local method agree on the clusterings quite well, whereas ICC often differs, mainly in granularity, as it tends to split the caves.



The clusters of a single cave in a 649-vertex graph; the fill color of the vertices indicates the clustering of ICC and the border color indicates that of the post-processed GMC; the local method selects the whole cave as a cluster.

Local clustering

Following and combining common criteria [10, 12], we want the clusters of a graph $G = (V, E)$ to be sets of vertices that are connected in G by many internal connections and only few connections outside, and achieve this by combining the local density with the relative density, both of which are computable from local information:

$$f(C) = \delta_\ell(C) \cdot \delta_r(C) = \frac{2 \text{deg}_{\text{int}}(C)^2}{|C|(|C|-1)(\text{deg}_{\text{int}}(C) + \text{deg}_{\text{out}}(C))}$$

Properties of the measure f :

- Avoids counterintuitive clusterings achieved by using either one of the two measures alone.
- Only relies on information obtainable from the adjacency lists of the included vertices.
- \Rightarrow A good approximation of the optimal cluster containing a given vertex can be obtained by local search [2].
- Local optimization is possible even for partially unknown graphs by online computation (cf. [9, 16]).
- \Rightarrow The method is likely to scale up for clustering graphs that are too large to be handled in the main memory, but have moderate neighborhood size.

The local search may either be complete, in which case the true optimal cluster is found, or heuristic, which often produces an approximate solution quickly. We have used simulated annealing [1] (similar approach to clustering of points in space has been taken in [17], simulated annealing for which is studied in [11]).

- Stochastically examine subsets of V containing v , and choose the candidate with maximal f as $\mathcal{C}(v)$.
- The initial cluster $\mathcal{C}'(v)$ of a vertex v contains v itself and all vertices adjacent to v .
- Each search step may either add a new vertex that is adjacent to an already included vertex, or remove an included vertex.
- Upon the removal of $u \in \mathcal{C}'(v)$, $u \neq v$, the connected component containing v becomes the next cluster candidate.

Using $\text{deg}_{\text{out}}(C) = |\{(u, v) \in E \mid u \in C, v \in V \setminus C\}|$ extends the above for directed graphs for which only the edges pointing out from the current vertex are locally accessible and incoming edges remain unknown until their source vertex is examined.

Applications

Due to its locality, the method is applicable for clustering e.g. the World-Wide Web with a crawler. Applications for Web clusters include, for example,

- grouping search results [3, 4, 8, 18],
- the identification of special interest groups or other types of web communities [7, 13]
- identification of link farms generated to fool PageRank,
- and the evaluation of generators of “Weblike” graphs [19] (cf. the evaluation of Internet topology generators in [14]).

Many of these applications generalize to other large natural networks as well. All vertices of a massive graph obviously cannot be used as starting points for clustering; hence a reliable method for uniform sampling of nonuniform natural networks would be of great interest.

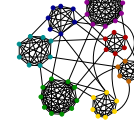
Discussing clustering in graphs does not exclude applications where the data consists of n -dimensional position vectors; several mappings of points in space into graphs are possible, using for example distance thresholds. Also textual data is easily mapped into a graph: similarities of DNA sequences can be built into a graph using the edit distances between pairs of sequences.

This research was supported by the Academy of Finland under grant 81120 and Helsinki Graduate School in Computer Science and Engineering (HCSE). We thank Marco Gaertler for providing the GMC and ICC clusterings and Kosti Rytönen for his help with the graph visualizations.

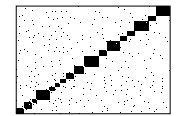
Experiments

We have conducted experiments on generation models for nonuniform random graphs, such as small-world [22] and scale-free [5] networks, as well as natural network data. In our experiments, the local search started at a randomly chosen vertex did not usually traverse the input graph extensively while locating the cluster of the start vertex. The overhead seems to depend on the graph structure: small-world networks seem “lighter” to search than scale-free or uniform random graphs of the same size and similar density [20].

For generalizations of the caveman graphs [21] consisting of a set of interconnected dense subgraphs of varying size [20], the method identifies “caves” as clusters regardless of the starting point of the search.



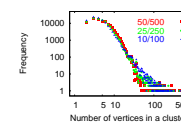
A caveman graph with 55 vertices and 217 edges; each cave (indicated by color) forms a cluster.



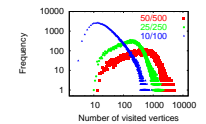
The adjacency matrix of a caveman graph of vertices and 1505 edges sorted by clusters of the local method.

We constructed a scientific collaboration network [15] from BibTeX data of mathematical publication databases [20]. Authors are the vertices of the network, identified by their first initial and surname. Authors are connected by an edge if they have a joint publication.

To study the effect of the heuristic search to the resulting clusters, we clustered the largest connected component of our collaboration graph ($|V| = 108,624$, $|E| = 333,546$), varying the number of independent restarts R per search vertex and the number of cluster modification steps S taken after each restart. We used three R/S -pairs, where $R \in \{10, 25, 50\}$ and $S = 10R$; the visit count drops as expected when R and S are decreased, but there is no difference of such magnitude in the distribution of the cluster sizes.



The distribution of the number of vertices per cluster for the collaboration graph.



The number of vertices visited while identifying a cluster for each vertex.

References

- [1] E. Aarts and J. Korst. *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Optimization and Neural Computing*. John Wiley & Sons, Inc., Chichester, UK, 1989.
- [2] E. Aarts and J. K. Lenstra, editors. *Local Search in Combinatorial Optimization*. John Wiley & Sons, Inc., Chichester, UK, 1997.
- [3] L. Adamic. The small world web. In S. Abiteboul and A.-M. Vercoustre, editors, *Proceedings of ECDL'99*, volume 1696 of *Lecture Notes in Computer Science*, pages 443–452, Berlin, Germany, 1999. Springer-Verlag.
- [4] G. Attardi, A. Gulli, and F. Sebastiani. Automatic Web page categorization by link and context analysis. In *Proceedings of 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence*, pages 105–119, 1999.
- [5] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, Oct. 1999.
- [6] U. Brandes, M. Gaertler, and D. Wagner. Experiments on graph clustering algorithms. In G. Di Battista and U. Zwick, editors, *Proceedings of the Eleventh European Symposium on Algorithms*, volume 2832 of *Lecture Notes in Computer Science*, pages 568–579, Berlin, Germany, Sept. 2003. Springer-Verlag, Heidelberg.
- [7] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of Web communities. In *Proceedings of the Sixth ACM SIGMOD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, New York, NY, USA, 2000. ACM Press.
- [8] E. J. Glover, K. Tsoutsoulikis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using Web structure for classifying and describing web pages. In *Proceedings of the 11th International World Wide Web Conference*, pages 562–569, New York, NY, USA, 2002. ACM Press.
- [9] S. Guha, N. Mishra, R. Motwani, and L. O’Callaghan. Clustering data streams. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pages 359–366, Los Alamitos, CA, USA, 2000. IEEE Computer Society Press.
- [10] R. Kannan, S. Vempala, and A. Vetta. On clusterings — good, bad and spectral. In *Proceedings of the 41st Annual Symposium on the Foundations of Computer Science*, pages 367–377, Los Alamitos, CA, USA, 2000. IEEE Computer Society Press.
- [11] R. W. Klein and R. C. Dubs. Experiments in projection and clustering by simulated annealing. *Pattern Recognition*, 22(2):213–220, Feb. 1989.
- [12] J. Kleinberg and S. Lawrence. The structure of the web. *Science*, 294(5548):1849–1850, Nov. 2001.
- [13] R. Kumar, A. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proceedings of the Twelfth International World Wide Web Conference*, pages 568–576, New York, NY, USA, 2003. ACM.
- [14] M. Mihail, C. Okunskiy, A. Saberi, and E. Zegura. On the semantics of internet topologies. Technical Report GIT-CC-02-07, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA, 2002.
- [15] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, USA*, 98(2):404–409, Jan. 2001.
- [16] L. O’Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani. Streaming data algorithms for high-quality clustering. In *Proceedings of 18th IEEE International Conference on Data Engineering*, pages 685–694, Los Alamitos, CA, USA, 2002. IEEE Computer Society Press.
- [17] J. Pacheco and O. Valencia. Design of hybrids for the minimum sum-of-squares clustering problem. *Computational Statistics & Data Analysis*, 43(2):235–248, June 2003.
- [18] A. Popescu, G. W. Flake, S. Lawrence, L. H. Ungar, and C. L. Giles. Clustering and identifying temporal trends in document databases. In *IEEE Advances in Digital Libraries*, pages 173–182, 2000.
- [19] S. Virtanen. Clustering the Chikan web. In *Proceedings of the First Latin American Web Congress*, pages 229–231, Los Alamitos, CA, USA, 2003. IEEE Computer Society Press.
- [20] S. Virtanen. Properties of nonuniform random graph models. Research Report A77, Helsinki University of Technology, Laboratory for Theoretical Computer Science, Espoo, Finland, May 2003.
- [21] D. J. Watts. *Small Worlds*. Princeton University Press, Princeton, NJ, USA, 1999.
- [22] D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small world’ networks. *Nature*, 393:440–442, June 1998.