Towards the Visualization of Overlapping Sets

Xavier Boyen
xb@cs.stanford.edu

Nina Mishra
nmishra@hpl.hp.com

Liadan O'Callaghan
loc@cs.stanford.edu*

Introduction Visualization is a key tool needed to evaluate the quality of data mining results. In this abstract we define and explore the problem of visualizing overlapping sets. We encountered this problem in the context of mining algorithms that produce overlapping clusters. We pose the following visualization problem, given a hypergraph $H = (V, E)$ with $|V| = n$ and $|E| = m$. Find a "layout" $(f; C_1, \ldots, C_m)$, where each C_i corresponds to a unique $e_i \in E$, and where $f : V \to \{1, \ldots, n\}^2$, such that: $C_1, \ldots, C_m$ are convex subsets of $\Re^2$; f is one-to-one; and for $1 \le i \le m$ C_i *respects* the hyperedges, i.e., $f(v) \in C_i \Leftrightarrow v \in e_i$. In what follows, for simplicity, the points in the range of f will sometimes be called "vertices", and if $v \in e_i$, the point $f(v)$ can be called a "member" of e_i or C_i without ambiguity.

Note that convexity by itself is not a sufficient requirement since the vertices can be placed arbitrarily around a circle and for each hyperedge $e_i \in E$, C_i can be the set whose border is the convex hull of the points representing the vertices of e_i. The problem with such a layout is that no proximity requirement is enforced: vertices in the same hyperedge may not be close to one another. We aim to find an f that respects the hyperedges and minimizes the total perimeter of the regions C_i[1]

Problem 1 *Given a hypergraph $H = (V, E)$, find an injective function $f : V \to \{1, \ldots, n\}^2$ and regions $C_1, \ldots, C_m$ that are convex subsets of $\Re^2$, such that (a) f respects the hyperedges in E, i.e., $f(v) \in C_i$ if and only if $v \in e_i$, and (b) the sum of the perimeters of the C_i is minimized.* [2]

This abstract describes results for linear and circular variants of the above planar problem.

*This work was done while the second two authors were at HP Research Labs, Palo Alto, CA. Second author partially supported by NSF Grant EIA-0137761. Third author partially supported by NSF Grants EIA-0137761 and IIS-0118173, and by an ARCS fellowship.

[1] An advantage of minimizing perimeter instead of, for example, area, is that "circle-like" regions are favored over "lanky" ones.

[2] We prevent f from mapping two vertices to arbitrarily close points by making the range of f a discrete set.

Linear Arrangements First, it is useful to consider embedding the hypergraph H in only one dimension, i.e., to find a mapping $f : V \to \{1, \ldots, n\}$ that minimizes total perimeter. In this case, the regions $C_1, \ldots, C_m$ in the optimal solution, and therefore, in all solutions that need to be considered, are intervals of $\Re$; perimeter, then, is just length. The following straightforward theorem can be proved via an algorithm for testing the incidence matrix of H for the "consecutive ones" property [BL76].

Theorem 2 *If $H = (V, E)$ is a hypergraph that has a mapping $f : V \to \{1, \ldots, n\}$ and regions $C_1, \ldots, C_m$ that respects the hyperedges in E, then a mapping f with minimum total perimeter can be found in time $O(|V| + |E|)$.*

Since very few hypergraphs admit linear layouts that respect the hyperedges, it is reasonable to relax the requirement that hyperedges be respected, and just minimize the total interval length. If the hypergraph is actually a graph, this relaxed problem, known as Optimal Linear Arrangement (OLA), is NP-hard [GJ79] but has an $O(\log n)$-approximation [RR98]. Treated as a hypergraph problem, OLA has an $O(d \log n)$-approximation, where d is the maximum degree, which is omitted due to space constraints.

Circular Arrangements It is also useful to consider mapping V to evenly-spaced points on the circle of circumference n, and representing each hyperedge not by a convex set, but by the shortest arc containing all of its members. Hyperedges may not be respected. Let the cost of laying out a hyperedge e on a circle be the arclength of the shortest arc containing the vertices in e. This cost is analogous to total interval length in the linear arrangement case. We call this problem "Optimal Circular Arrangement" (OCA). Note that $\sum_{e \in E} |e|$ is a lower bound on the cost of a solution to OCA. This bound is attained if the hyperedges are respected, that is, if the arc for each hyperedge contains exactly the member vertices. For hypergraphs that admit such a layout, Booth and

Leuker's linear-time "circular ones" algorithm [BL76] finds the corresponding layout. Most hypergraphs are unlikely to have the "circular ones" property, but the OCA objective function is appropriate in that minimizing this objective will tend to produce layouts that respect hyperedges as much as possible.

Another objective function that encourages a layout to respect hyperedges is the sum, over all hyperedges, of the Euclidean length of the longest edge of the convex hull of the member vertices. In a graph, this quantity is the sum of Euclidean lengths of the edges, when each edge is drawn as the chord joining its endpoints. We will call the problem of minimizing this second function, whether for graphs or for hypergraphs, Optimal Chordal Arrangement (OCHA). A reduction from OLA shows that OCA and OCHA are NP-hard, even in graphs.[3] Liberatore proved OCA NP-hard [Lib02].

Definition 3 *Let $G = (V, E)$ be a graph, and let $f : V \to \{1, \ldots, n\}$ be one-to-one. Then $C_L(f, G)$ is the cost of f as an OLA solution for G, and $C_C(f, G)$ is the cost of f as an OCA solution for G. That is, $C_L(f, G) = \sum_{(u,v)\in E} |f(u) - f(v)|$ and $C_C(f, G) = \sum_{(u,v)\in E} \text{ARCLENGTH}(f(u), f(v))$. Also, $C_{Ch}(f, G)$ is the cost of f as an OCHA solution for G; $C_{Ch}(f, G) = \sum_{(u,v)\in E} \text{CHORDLENGTH}(f(u), f(v))$, where $\text{CHORDLENGTH}(a, b)$ is the length of the chord joining the endpoints of the arc whose arclength is $\text{ARCLENGTH}(a, b)$.*

Note that for every OLA solution g there is a corresponding solution g' to either OCA or OCHA, in which the vertices mapped by g to 1, 2, ..., n are assigned in clockwise order to evenly-spaced points around a circle of perimeter n. Furthermore, $C_C(g', G) \le C_L(g, G)$, since edges that were longer than $n/2$ under g will cost less under g'. Since chordal length is less than arclength, $C_{Ch}(g', G) \le C_C(g, G)$ and thus $C_{Ch}(g', G) \le C_L(g, G)$. We show that g' is in fact a good chordal arrangement.

Theorem 4 *For every graph G and every OCHA solution f for G there is an OLA solution f' with $C_L(f', G) \le \pi C_{Ch}(f, G)$.*

Proof: Suppose the evenly-spaced points to which f assigns the members of V are $x_1, \ldots, x_n$ in clockwise order. Converting f to an OLA solution just requires labelling some x_k '1,' and numbering the other points 2, 3, ..., n in clockwise order. This conversion corresponds to making a "cut" between x_k and x_{k-1}. Consider a particular cut. For each $e \in E$, let s_e denote the length of e under f, and let l_e denote its length after the cut. Consider first the edges (x_i, x_j) with $i < k$ and $j - i < n/2$ (i.e., edges (x_i, x_j) such that the cut occurred somewhere along the shortest arc from x_i to x_j); these edges will be said to "cross the cut." If $e \in E$ crosses the cut, then $l_e - s_e \le n$, since $l_e \le n$. If $e \in E$ does not cross the cut, $l_e - s_e \le (\frac{\pi}{2} - 1)s_e$, since l_e will be the length of the arc joining the endpoints of e, under f.

There are n possible cuts (i.e., conversions to OLA solutions). Fix $e \in E$. e crosses the cut in at most $(\pi/2)s_e$ conversions, since the number of cuts that e crosses is the length of the arc joining the endpoints of e under f. Clearly there are at most n cuts that e does *not* cross. The sum, over all n conversions, of the increase in the cost of e due to the conversion, is then at most $(\frac{\pi}{2}s_e)n + n(\pi/2 - 1)s_e \le (\pi - 1)ns_e$. The sum over all edges e and over all conversions, of the increase in the cost of e due to the conversion, is at most $\sum_{e\in E}(\pi - 1)ns_e$. The average conversion, then, adds at most $\sum_{e\in E}(\pi - 1)s_e$ to the cost; therefore there must exist a conversion that produces a linear arrangement f' with $C_L(f', G) \le (\pi - 1 + 1)C_{Ch}(f, G) = \pi C_{Ch}(f, G)$. ∎

A nearly identical proof shows that for every OCA solution there is an OLA solution of at most twice the cost. The following is immediate.

Theorem 5 *If OLA is α-approximable, then OCA is approximable within 2α and OCHA within $\pi\alpha$.*

The best known α is $O(\log n)$ [RR98], so OCA has an $O(\log n)$-approximation; Liberatore [Lib02] proved that OCA was $O(\log n)$-approximable, but the above proof is simpler.

References

[BL76] Booth and Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using pq-tree algorithms. *JCSS*, 13:335–379, 1976.

[GJ79] Garey and Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W.H. Freeman, New York, 1979.

[Lib02] V. Liberatore. Circular arrangements. In *Proc ICALP*, 2002.

[RR98] Rao and Richa. New approximation techniques for some ordering problems. In *SODA*, 1998.

[3] The proof introduces enough extra dummy vertices so that the best circular arrangement is the best linear arrangement.