Faster Pathfinder algorithm for sparse networks

Vladimir Batagelj, Anže Vavpetič

University of Ljubljana

Sunbelt XXX, Riva del Garda, Italy, June 29 - July 4, 2010

Outline

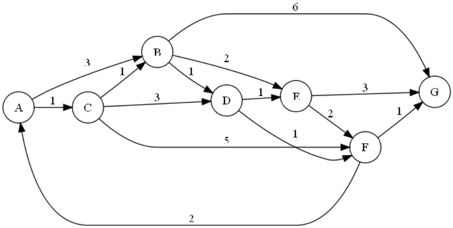
Pathfinder

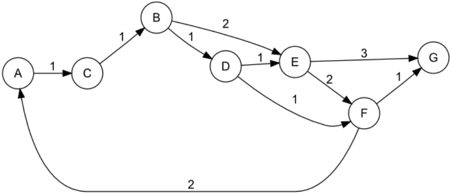
The Pathfinder algorithm was proposed in eighties (Schvaneveldt 1981, Schvaneveldt etal. 1989; Schvaneveldt, 1990) [14, 13, 15] for simplification of weighted networks – it removes from the network all lines that do not satisfy the triangle inequality – if for a line a shorter path exists connecting its endpoints then the line is removed. The basic idea of the Pathfinder algorithm is simple. It produces a network $PFnet(\mathbf{W}, r, q) = (\mathcal{V}, \mathcal{L}_{PF})$

compute $\mathbf{W}^{(q)}$;
$\mathcal{L}_{PF} := \emptyset$;
for $e(u, v) \in \mathcal{L}$ **do begin**
  if $\mathbf{W}^{(q)}[u, v] = \mathbf{W}[u, v]$ **then** $\mathcal{L}_{PF} := \mathcal{L}_{PF} \cup \{e\}$
end;

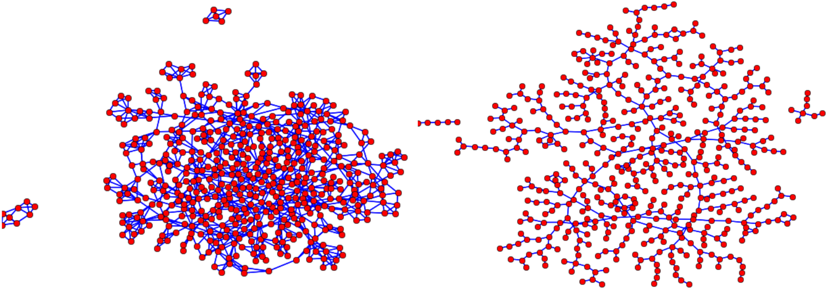
where $\mathbf{W}$ is a network dissimilarity matrix and $\mathbf{W}^{(q)}$ the matrix of values of all walks of length at most q computed over the semiring $(\mathbb{R}_0^+, \oplus, \boxed{r}, \infty, 0)$ with $a \boxed{r}\, b = \sqrt[r]{a^r + b^r}$ and $a \oplus b = \min(a, b)$.

Pathfinder

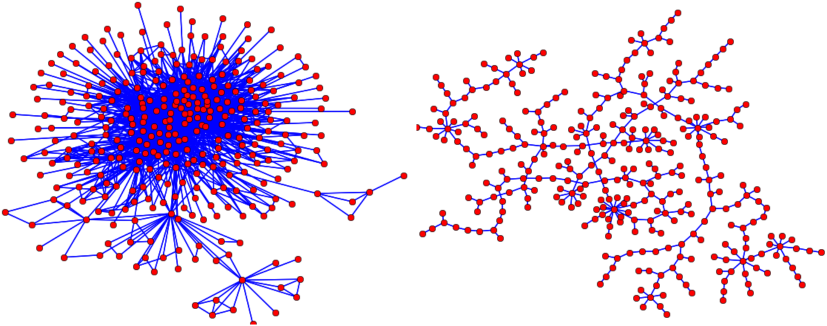




Pathfinder



Pathfinder



Pathfinder – theoretical results 1

Theoretical results [7]: For a given dissimilarity matrix **W** the $PFnet(\mathbf{W}, r, q)$

- Is unique
- Preserves geodetic distances
- Links nearest neighbors
- Contains the same information as the minimum method of hierarchical clustering
- $PFnet(\mathbf{W}, r = \infty, q = n - 1)$ is the union of all MINTREES

Pathfinder – theoretical results 2

It holds also:

- Graph $PFnet(\mathbf{W}, r_2, q)$ is a spanning subgraph of graph $PFnet(\mathbf{W}, r_1, q)$ iff $r_1 < r_2$
- Graph $PFnet(\mathbf{W}, r, q_2)$ is a spanning subgraph of graph $PFnet(\mathbf{W}, r, q_1)$ iff $q_1 < q_2$
- Similarity transformations preserve structure: the graph $PFnet(\mathbf{W}, r, q)$ is equal to the graph $PFnet(\alpha\mathbf{W}, r, q)$ for $\alpha > 0$.
- Monotonic transformations preserve structure for all $r = \infty$: the graph $PFnet(\mathbf{W}, r = \infty, q)$ is equal to the graph $PFnet(f(\mathbf{W}), r = \infty, q)$ where $f : \mathbb{R} \to \mathbb{R}$ is strictly increasing mapping and $f(\mathbf{W} = [f(w_{ij})]$.

Pathfinder – the original algorithm

In the original algorithm the matrix $\mathbf{W}^{(q)}$ is computed on the basis of its definition

$$\mathbf{W}^{(q)} = \sum_{i=0}^{q} \mathbf{W}^i$$

by computing all its powers $\mathbf{W}^i$, $i = 1, \ldots, q$. The complexity of the algorithm is $O(qn^3)$, therefore $O(n^4)$, for $q \geq n - 1$. Therefore it can be applied only to relatively small (up to some hundreds vertices) networks.
Interest for Pathfinder transformation was renewed around the year 2000 by Chen [5].

Semirings – Computing the closure over a semiring with absorption

Because in our case the set of vertices $\mathcal{V}$ is finite, so is the set of all paths $\mathcal{E}_{uv}$. Therefore we can compute the value of all walks $w(\mathcal{S}^\star_{uv}) = w(\mathcal{E}_{uv})$. One possibility is to use for large enough k the equality:

$$\mathbf{W}^\star = \mathbf{W}^{(k)} = (\mathbf{1} + \mathbf{W})^k$$

To speed-up the computation we can consider the sequence $(\mathbf{1} + \mathbf{W})^{2^i}, i = 1, .., s$.
It turned out that this is not the fastest way to compute the $\mathbf{W}^\star$.

Semirings – Computing the closure over a complete semiring

Kleene, Warshall, Floyd and Roy contributed to the development of the procedure which final form was given by Fletcher [6].

$\mathbf{C}_0 := \mathbf{W}$;
for $k := 1$ **to** n **do begin**
  for $i := 1$ **to** n **do for** $j := 1$ **to** n **do**
    $c_k[i,j] := c_{k-1}[i,j] + c_{k-1}[i,k] \cdot (c_{k-1}[k,k])^\star \cdot c_{k-1}[k,j]$;
  $c_k[k,k] := 1 + c_k[k,k]$;
end;
$\mathbf{W}^\star := \mathbf{C}_n$;

If we delete the statement $c_k[k,k] := 1 + c_k[k,k]$ we obtain the algorithm for computing the strict closure $\overline{\mathbf{W}} = \mathbf{W}\mathbf{W}^\star$.

Semirings – Dissimilarities

Joly and Le Calvé theorem [8]:
For any even dissimilarity measure d there is a unique number $p \geq 0$, called its *metric index*, such that: d^r is metric for all $r \leq p$, and d^r is not metric for all $r > p$.

In the opposite direction we can say: Let d be a dissimilarity and for x, y and z we have $d(x,z) + d(z,y) \geq d(x,y)$ and $d(x,y) > \max(d(x,z), d(z,y))$ then there exists a unique number $p \geq 0$ such that for all $r > p$

$$d^r(x,z) + d^r(z,y) < d^r(x,y)$$

or equivalently

$$d(x,z) \boxed{r}\, d(z,y) < d(x,y)$$

Semirings – Minkowski operation

Minkowski operation $a \boxed{r} b = \sqrt[r]{a^r + b^r}$:

$r = 1 \Rightarrow a \boxed{r} b = a + b$,
$r = 2 \Rightarrow a \boxed{r} b = \sqrt{a^2 + b^2}$,
$r = \infty \Rightarrow a \boxed{r} b = \max(a, b)$.

And let $a \oplus b = \min(a, b)$.
The structure $(\mathbb{R}_0^+, \oplus, \boxed{r}, \infty, 0)$ is a complete semiring with $a^\star = 0$. It is called also *Pathfinder* semiring.

Spanish algorithms

Since the Pathfinder semiring is idempotent it holds

$$\mathbf{W}^{(q)} = (\mathbf{1} \oplus \mathbf{W})^q$$

This power can be computed faster using binary algorithm (for example, to compute $a^{57} = a^{32} \cdot a^{16} \cdot a^{8} \cdot a^{1}$ we need only 8 multiplications instead of 56). This improvement was proposed by Guerrero-Bote etal. (2006) [7] and reduces complexity to $O(n^3 \log q)$. When $q \geq n-1$, $\mathbf{W}^{(q)} = \mathbf{W}^\star$ and it can be determined by the Fletcher's algorithm over Pathfinder semiring. This improvement was proposed by Quirin etal. (2008) [9] and reduces complexity to $O(n^3)$. Additional improvement can be made for undirected networks in the case $q \geq n-1$ and $r = \infty$. In this case the network PF is the union of all minimal spanning trees of N. It can be obtained using an adapted version of Kruskal's minimal spanning tree algorithm as described in Quirin etal. (2008) [10]. The complexity of this algorithm is $O(m \log n)$ where m is the number of edges.

Sparse Pathfinder

For sparse networks in general case there is still some space for improvements. We rewrite the basic Pathfinder algorithm in the form

$\mathcal{L}_{PF} := \emptyset$;
for $v \in \mathcal{V}$ **do begin**
  compute the list $S = ((u, d_u) : u \in N(v))$, where $d_u = \mathbf{W}^{(q)}[v, u]$;
  for $(u, d_u) \in S$ **do**
    if $d_u = \mathbf{W}[v, u]$ **then** $\mathcal{L}_{PF} := \mathcal{L}_{PF} \cup \{(v, u)\}$
end;

$N(v)$ denotes the set of successors of vertex v.
For determining the values $d_u = \mathbf{W}^{(q)}[v, u]$ for $q = n - 1$ we can use an adapted Dijkstra's algorithm that determines the list S in a single run. The job is done when all values of vertices from $N(v)$ are determined. Only a (small) portion of network should be inspected for each vertex v. To efficiently implement this algorithm a special data structure *Indexed Priority Queue* is needed.

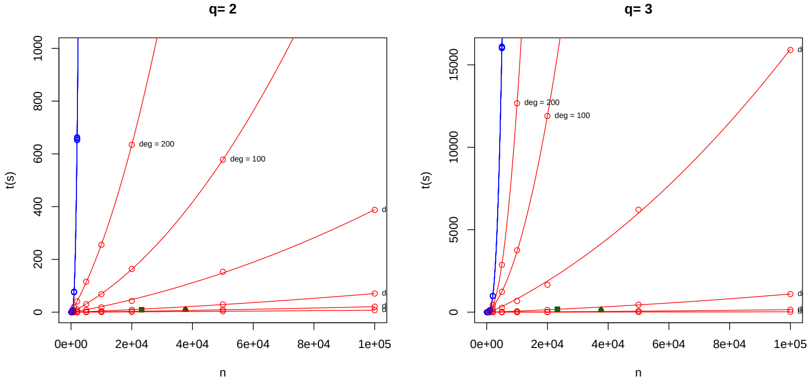
Sparse Pathfinder – BFS algorithm

In the case $q < n - 1$ a variant of BFS (Breath First Search) algorithm is used to determine the list S.
The FIFO queue Q is composed of triples (t, d, l): t is a vertex, d is a dist-length and l is a line-length.
To make the implementation fast all the structures: the queue Q and lists *Plist* and *Vlist* are represented with arrays.

Sparse Pathfinder – BFS algorithm compute the list S

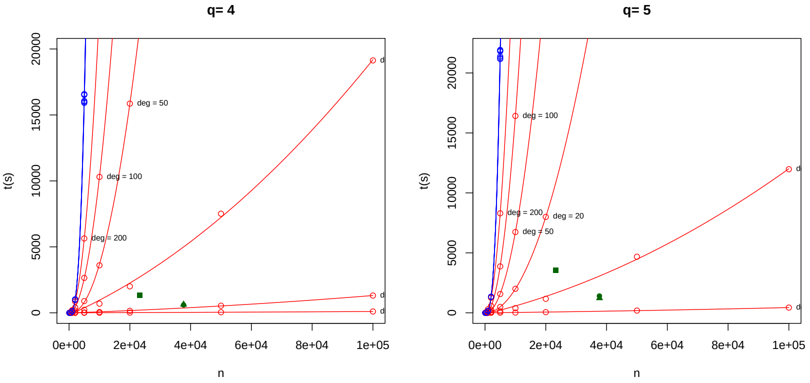

```
S := ∅; T := N(v); emptyQ; dMax := max{w_vu : u ∈ T};
putLastQ(v, 0, 0); dist[v] := 0; level := 0;
while sizeQ() > 0 do begin
    (u, d_u, l) := firstFromQ(); l := l + 1;
    if l > level then begin
        level := l;
        for v ∈ Plist do P[v] := 0;
        nPlist := 0;
    end
    for t ∈ N(u) do begin
        dNew := d_u ⊡ʳ w(u, t);
        if dNew ≤ dMax then begin
            if V[t] then begin
                if dNew < dist[t] then begin
                    dist[t] := dNew;
                    if l < q then begin
                        if P[t] > 0 then updateQ(t, dNew)
                        else putLastQ(t, dNew, l);
                    end
                end
            end else begin
                dist[t] := dNew; if l < q then putLastQ(t, dNew, l);
            end
        end
    end
end;
for v ∈ Plist do P[v] := 0; for v ∈ Vlist do V[v] := false;
nPlist := 0; nVlist := 0;
for t ∈ T do S := S ∪ {(t, dist[t])};
```

Tests – $q = 2$ and $q = 3$

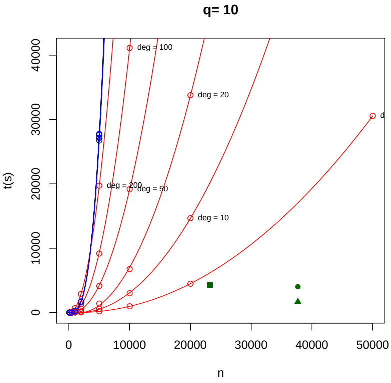


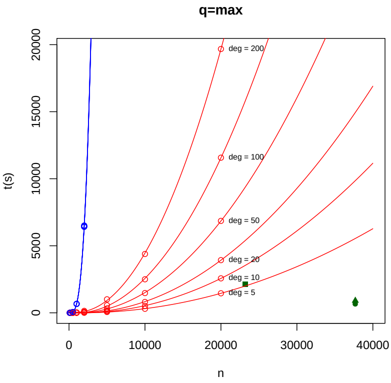
eatRSd5.net: $n = 23219$, $\overline{\text{deg}} = 28.048$ Edinbourgh Associtive Thesaurus, d_5
Cluster1.net: $n = 37689$, $\overline{\text{deg}} = 15.875$ Citations in Clustering
$d(u, v) = 1 - n(u, v)/\max(inS(u), inS(v), outS(u), outS(v))$
Cluster2.net: $n = 37690$, $\overline{\text{deg}} = 16.016$ Citations in Clustering
$d(u, v) = 1/n(u, v)$

Tests – $q = 4$ and $q = 5$



Tests – $q = 10$ and $q = max$





Conclusions

- the tests with sparse random networks of Erdos-Renyi type show that the new algorithms extend the range of sparse networks for which we can determine the Pathfinder network in reasonable time to at least $n = 50000$.
- it seems that on real-life networks (green marks) the algorithm works much faster than on random networks with the same average degree.

References I

- A. V. Aho, J. E. Hopcroft, J. D. Ullman, *The Design and Analysis of Computer Algorithms.* Addison-Wesley, Reading, Massachusetts (1976).
- Batagelj, V.: *Semirings for Social Networks Analysis*. Journal of Mathematical Sociology, **19**(1994)1, 53-68.
- R.E. Burkard, R.A. Cuninghame-Greene, U. Zimmermann, eds., *Algebraic and Combinatorial Methods in Operations Research.* Annals of Discrete Mathematics **19** (1984).
- B. Carré, *Graphs and Networks.* Clarendon, Oxford (1979).
- Chen, Chaomei: Generalised similarity analysis and Pathfinder network scaling. Interacting with Computers, 10 (2): pp. 107-128.
- J. G. Fletcher, "A more general algorithm for computing closed semiring costs between vertices of a directed graph," *CACM* (1980), pp. 350-351.

References II

Vicente P. Guerrero-Bote, Felipe Zapico-Alonso, Mara Eugenia Espinosa-Calvo, Roco Gómez Crisóstomo, Flix de Moya-Anegón: Binary Pathfinder: An improvement to the Pathfinder algorithm. Information Processing and Management, Volume 42, Issue 6, December 2006, Pages 1484-1490. http://linkinghub.elsevier.com/retrieve/pii/S0306457306000367

Joly S., Le Calvé G. (1986) Etude des puissances d'une distance. Statistique et Analyse de Données, 11/3, 30-50.

A. Quirin, O. Cordón, J. Santamaria, B. Vargas-Quesada, F. Moya-Anegón: A new variant of the Pathfinder algorithm to generate large visual science maps in cubic time. http://www.scimago.es/benjamin/A_new_variant_of_the_Pathfinder_Algorithm.pd

Arnaud Quirin, Oscar Cordón, Vicente P. Guerrero-Bote, Benjamn Vargas-Quesada and Felix Moya-Anegón: A Quick MST-Based Algorithm to Obtain Pathfinder Networks $(\infty, n-1)$. Journal of the American Society for Information Science and Technology, Volume 59, Issue 12 (p 1912-1924). http://www3.interscience.wiley.com/cgi-bin/fulltext/120736756/PDFSTART

References III

- F. S. Roberts, *Discrete Mathematical Models.* Prentice-Hall, Englewood Cliffs, New Jersey (1976). Amazon.
- Schvaneveldt, R. W., Dearholt, D. W., Durso, F. T. (1988) Graph theoretic foundations of Pathfinder networks. Comput. Math. Applic. **15** (4), 337-345.
- Schvaneveldt, R. W. (Ed.) (1990) Pathfinder Associative Networks: Studies in Knowledge Organization. Norwood, NJ: Ablex. http://interlinkinc.net/PFBook.zip
- Schvaneveldt, R. W., Durso, F. T., Dearholt, D. W. (1989) Network structures in proximity data. In G. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory, Vol. 24 (pp. 249-284). New York: Academic Press. http://www.interlinkinc.net/Roger/Papers/Schvaneveldt_Durso_Dearholt_1989.p
- Interlink – Tools for Pathfinder Network Analysis. http://www.interlinkinc.net/
- U. Zimmermann, *Linear and Combinatorial Optimization in Ordered Algebraic Structures.* Annals of Discrete Mathematics **10** (1981).