

Algorithms for Biological Graphs:
Analysis and Enumeration
ICTCS Doctoral Research Awards
15th Italian Conference on Theoretical Computer Sciences

Andrea Marino
Tutor: Prof. Pierluigi Crescenzi

Dipartimento di Informatica, University of Milan

18th September 2014

Modelling biological interactions through graphs.

- **Biological networks modeled through simple graphs** neglects hyper connections and arc dependencies.
- **Biological networks are potential networks** not all the edges or nodes exist at the same time.

- **When looking for structures in these graphs**, not all of them make sense.
- **What Biologists want:** seeing all the candidate structures and select a posteriori the good ones.



Klein Cecilia, Marino Andrea, Sagot Marie-France, Vieira-Milreu Paulo, Brilli Matteo: Structural and dynamical analysis of biological networks. Briefings in Functional Genomics

The number of objects to be enumerated could be exponential.

Polynomial Total Time... *The least that we could ask is that the time required to output all solutions be bounded by a polynomial in n [the size of the input] and C [the number of solutions]. ...*

Polynomial Delay... *The delay between any two consecutive solutions, is bounded by a polynomial in the input size. ...*

in *On generating all maximal independent sets,*
Johnson, Yannakakis, Papadimitriou,

IPL 1988.

Part I

Enumerating Central or Peripheral Node

Which one are the most and least important entities in the biological network for the organism's viability?

The peripheral (central) nodes have often maximum (minimum) eccentricity.

Definition (Distance, Eccentricity, Radius, and Diameter)

In an unweighted undirected graph $G = (V, E)$ connected.

- The distance $d(u, v)$, number of edges along shortest path from u to v .
- The eccentricity of a node u , $\text{ecc}(u) = \max_{v \in V} d(u, v)$.
- The diameter $D = \max_{u, v \in V} d(u, v) = \max_{u \in V} \text{ecc}(u)$
- The radius $R = \min_{u \in V} \text{ecc}(u)$

They can be extended for directed weighted (strongly connected) graphs.

We will focus on Diameter and nodes with max eccentricity.

Enumerating Nodes with maximum Eccentricity

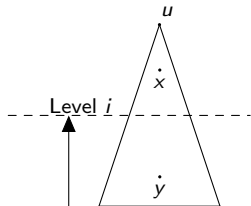
D and all the nodes having max eccentricity can be computed by doing n BFSes, i.e. computing the eccentricity of each node one after the other.

Our Algorithm is like textbook algorithm, but:

- Specify the **ORDER** in which BFSes have to be executed.
- Refine a lower bound lb of D , i.e. the maximum ecc found until that moment with the nodes with that ecc.
- Refine an upper bound ub for the ecc of the remaining nodes.
- Stop when the remaining nodes cannot have eccentricity higher than the lower bound we have.

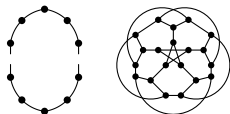
ORDER: Pick a node u (possibly high degree), consider nodes in decreasing distance from u .

- The eccentricity of the nodes x is bounded by $\max\{\max_y \text{ecc}(y), 2(i-1)\} = \max\{lb, 2(i-1)\}$



Experiments


- In the worst case we perform n visits like in the traditional algorithm.
 - When all the nodes have the same eccentricity.



- In almost all Metabolic Net. number of visits of our algorithm is $<10\%n$ and in almost all Protein Interaction Net., visits $<1\%n$.
- In the great majority of real world net., visits is less than 100-1000, even for networks with hundreds of millions of nodes.
- In the Facebook graph (68.7G edges) the number of visits to find the diameter (just a diametral pair) is 17 instead of 721.1 millions.

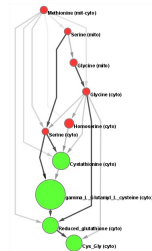
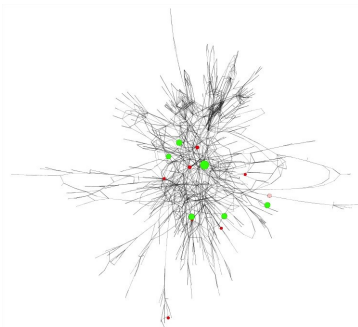
-  Crescenzi, P., Grossi, R., Habib, M., Lanzi, L., Marino, A.: On Computing the Diameter of Real-World Directed (Weighted) Graphs. SEA 2012.
-  Crescenzi, P., Grossi, R., Habib, M., Lanzi, L., Marino, A.: On Computing the Diameter of Real-World Undirected Graphs. Workshop on Graph Algorithms and Applications (Zurich July 3, 2011). Theoretical Computer Science 2012.
-  Crescenzi, P., Grossi, R., Imbrenda, C., Lanzi, L., Marino, A.: Finding the Diameter in Real-World Graphs: Experimentally Turning a Lower Bound into an Upper Bound. ESA 2010.

Recently improved:

-  Borassi M., Crescenzi P., Habib M., Kusters W.A., Marino A., Takes F.W.: On the Solvability of the Six Degrees of Kevin Bacon Game - A Faster Graph Diameter and Radius Computation Method. FUN 2014.

Part II

Enumerating Stories



In a metabolic network, given a set of interesting compounds (called black nodes) a story is the cascade of all reactions connecting them.

Definition (Story and Minimal Story Arc Set (SAS))

- Given a directed graph with nodes black and white, a **story** is a maximal acyclic subgraph in which just black nodes can have in degree or out degree 0.
- A Minimal SAS is the set of arcs to be removed to get a **story**.

Definition (**In the literature.** Minimal Feedback Arc Set (FAS))


Arcs to be removed to get a maximal acyclic sub graph.


An efficient minimal FAS enumerator exists but we cannot run it to get stories.

- A minimal FAS is not necessarily a minimal SAS
 - there could be white nodes with in or out degree zero.
- A minimal SAS is not necessarily a minimal FAS if there are bad nodes.
 - A **bad node** is a white node v such that any pair of arcs (u, v) and (v, w) is part of a cycle.


Brute force algorithm able to enumerate all the stories:

- by inspecting all the orderings of the nodes, considering arcs consistent with the orderings, trying to make the graph a story.
- check whether already generated before to output.

 V. Acuña, E. Birmelé, L. Cottret, P. Crescenzi, F. Jourdan, V. Lacroix, A. Marchetti-Spaccamela, A. Marino, P.V. Milreu, M.-F.Sagot, L. Stougie. Telling stories. Presented at Workshop on Graph Algorithms and Applications (Zurich, July 3, 2011) and published by Theoretical Computer Science, 2012.

 V. Acuña, E. Birmelé, L. Cottret, P. Crescenzi, F. Jourdan, V. Lacroix, A. Marchetti-Spaccamela, A. Marino, P.V. Milreu, M.-F.Sagot, L. Stougie. Metabolic stories: uncovering all possible scenarios for interpreting metabolomics data. In First RECOMB Satellite Conference on Open Problems in Algorithmic Biology (RECOMB-AB), August 27-29, 2012, St. Petersburg, Russia.

Recently improved in the following

 M. Borassi, P. Crescenzi, V. Lacroix, A. Marino, P.V. Milreu, M.-F.Sagot. Telling Stories Fast Via Linear-Time Delay Pitch Enumeration. In SEA 2013.

Open Question: What is the complexity of enumerating stories?

Part III

Enumerating Cycles

Paths and cycles in Biological Interaction Networks show the different routes along which a molecule can affect another.

Listing cycles is equivalent to listing paths: we focus on st-paths.

- An optimal algorithm to list paths should take $O(n + m + L)$ time to enumerate all simple cycles, where L denotes the sum of the lengths of all the paths,
- There is an infinite class of biconnected undirected graphs for which Johnson algorithm (1975), i.e. the best known algorithm to enumerate paths or cycles, is not optimal.

$\mathcal{P}(G)$ is the set of all paths in G , $\mathcal{P}_{s,t}(G)$ is the set of all st -paths in G .

Recursively apply the binary partition method: dividing $\mathcal{P}_{s,t}(G)$ in

- 1 **Left Branch:** the set of paths that *use* an edge $e = (s, v)$,
- 2 **Right Branch:** the set of paths that do *not* use edge e

$$\mathcal{P}_{s,t}(G) = \underbrace{(s, v) \cdot \mathcal{P}_{v,t}(G - s)}_{\text{left-branch}} \cup \underbrace{\mathcal{P}_{s,t}(G - (s, v))}_{\text{right-branch}}$$

Goal 1: there always exists at least one st -path in each recursive call;

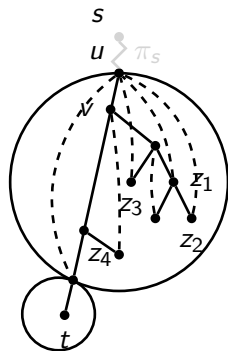
Goal 2: the amortized cost of finding a path π should be $O(|\pi|)$ in the overall time complexity.

Maintaining a certificate in constant amortized cost

Definition (Certificate)

During the recursion at u , a *certificate* C is a suitable data structure that uses a DFS tree rooted at u to classify its edges as tree edges or back edges.

When doing the recursion at u , the DFS is going to change just for the biconnected component of u .



- The cost of an update of the certificate is proportional to the number of solutions produced because of that update.
- Depending on the edge we choose, we know a priori whether a left or right branch is productive or not.



Etienne Birmelé, Rui Ferreira, Roberto Grossi, Andrea Marino, Nadia Pisanti, Romeo Rizzi, Gustavo A.T. Sacomoto. Optimal Listing of Cycles and **st**-Paths in Undirected Graphs. SODA 2013.

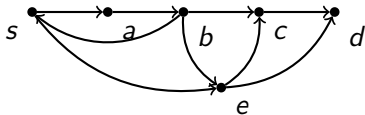
Open Question: Is it possible to do the same in the case of directed graphs?

Part IV

Enumerating Bubbles

The different alternative ways of generating mRNA from DNA (*alternative splicing*) correspond to the pairs of paths in a graph representing the reads.

Given two vertices s and t in a directed graph, an (s, t) -path is a path from s to t . An (s, t) -bubble is a pair of two vertex-disjoint (s, t) -paths.



$\langle s, e \rangle$ and $\langle s, a, b, e \rangle$ form an (s, e) -bubble.

Turning bubbles starting from s into special cycles

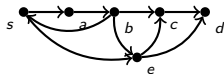
Given s , we transform G into a new graph $G'_s = (V'_s, E'_s)$.

Definition

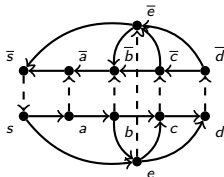
A *bubble-cycle* in G'_s is a cycle containing no pair of twins except for (s, \bar{s}) and (t, \bar{t}) .

Lemma

Correspondence between the (s, t) -bubbles in G , and the bubble-cycles from s .



(a) Graph G



(b) Graph G'_s

$\langle s, e, \bar{e}, \bar{b}, \bar{a}, \bar{s}, s \rangle$ is a bubble-cycle: it corresponds to an (s, e) -bubble composed by $\langle s, e \rangle$ and $\langle s, a, b, e \rangle$.

The Johnson Algorithm is the best known algorithm to enumerate cycles in directed graphs, recursively traversing the graph (backtrack)

Adapting the Johnson Algorithm to enumerate bubble-cycles

taking care of:

- exploring just vertices not producing too many twins.
- **Not Trivial:** maintaining the same time complexity, i.e. linear time delay.



Etienne Birmelé, Pierluigi Crescenzi, Rui Ferreira, Roberto Grossi, Vincent Lacroix, Andrea Marino, Nadia Pisanti, Gustavo A.T. Sacomoto, Marie-France Sagot. Enumerating bubbles in directed graphs with linear delay. SPIRE 2012

Open Question: What about enumerating triples, quadruplets of paths?

Thanks