Linear arrangement of vertices

Ramon Ferrer-i-Cancho & Argimiro Arratia

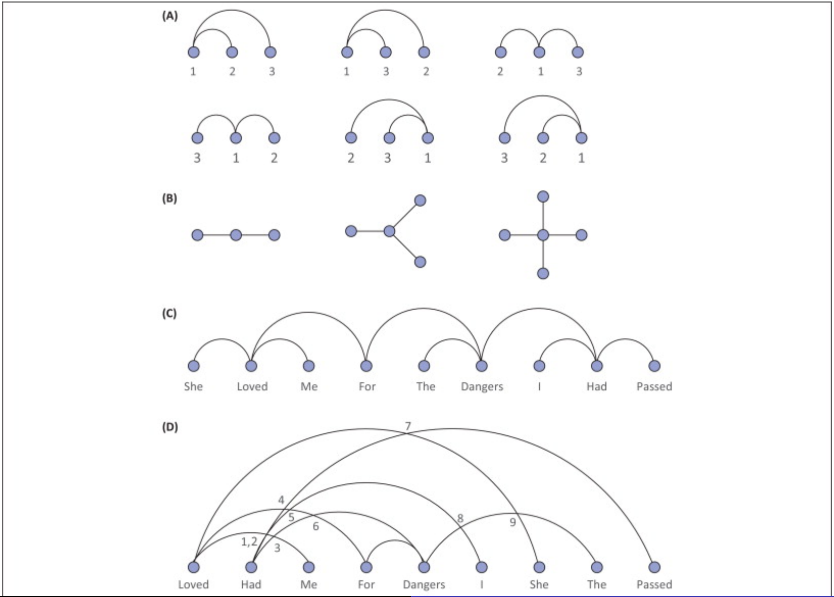
Universitat Politècnica de Catalunya

Version 0.4
Complex and Social Networks (2016-2017)
Master in Innovation and Research in Informatics (MIRI)

Official website: `www.cs.upc.edu/~csn/`

Contact:

- Ramon Ferrer-i-Cancho, rferrericancho@cs.upc.edu, `http://www.cs.upc.edu/~rferrericancho/`
- Argimiro Arratia, argimiro@cs.upc.edu, `http://www.cs.upc.edu/~argimiro/`


(A)
1 2 3
1 3 2
2 1 3
3 1 2
2 3 1
3 2 1
(B)
(C)
She Loved Me For The Dangers I Had Passed
(D)
7
4
5
6
8
9
1,2
3
Loved Had Me For Dangers I She The Passed

Two interesting properties:

- The linear (euclidean) distance between connected words is "small".
- The number of crossings is "small".

An statistical challenge:

- Are they significantly small?
- What would be a suitable null hypothesis?

A scientific question: if they are significantly small, then why?

Focus on trees

A linear arrangement of vertices

- Vertices are labelled with numbers $1, 2, 3, ..., n$ being n the number of vertices of the network.
- $s, t, u, v, ...$ designate vertices.
- A linear arrangement of vertices is one of the $n!$ possible orderings of n vertices.
- A linear arrangement can be defined by $\pi(v)$, the position of vertex v in the ordering ($\pi(v) = 1$ if v is the first vertex, $\pi(v) = 2$ if v is the second vertex and so on...).
- For a linear arrangement of a tree, the mean edge length is defined as

$$\langle d \rangle = \frac{D}{n-1} = \frac{1}{n-1} \sum_{u \sim v} |\pi(u) - \pi(v)| \tag{1}$$

Edge crossings

Two edges $u \sim v$ and $s \sim t$ such that $\pi(u) < \pi(v)$ and $\pi(s) < \pi(t)$ cross if and only if

- $\pi(u) < \pi(s) < \pi(v) < \pi(t)$ or
- $\pi(s) < \pi(u) < \pi(t) < \pi(v)$

Example with 4 vertices.
The number of crossings is

$$C = \frac{1}{2} \sum_{u \sim v} C_{u,v}, \tag{2}$$

where $C_{u,v}$ is the number of edge crossings involving $u \sim v$.
$C \geq 0$, but what is the maximum value of C?

Degrees in trees

- Mean degree is constant, i.e.

$$\langle k \rangle = \frac{1}{n} \sum_{v=1}^{n} k_v = 2 - 2/n. \quad (3)$$

- Degree variance is fully determined by the 2nd moment, i.e.

$$V[k] = \left\langle k^2 \right\rangle - \langle k \rangle^2 = \left\langle k^2 \right\rangle - (2 - 2/n)^2 \quad (4)$$

- The 2nd moment is minimized by a linear tree and maximized by a star tree, i.e. [Ferrer-i-Cancho, 2013]

$$\text{(linear tree) } 4 - \frac{6}{n} \leq \left\langle k^2 \right\rangle \leq n - 1 \text{ (star tree)} \quad (5)$$

Mean edge length in trees

- Real syntactic dependency trees: sublinear growth (Fig. of [Ferrer-i-Cancho, 2004]).
- Some theoretical bounds [Ferrer-i-Cancho, 2013]
  - In a random linear arrangement, $E[\langle d \rangle] = \frac{n+1}{3}$.
  - In a non-crossing tree, $\langle d \rangle \leq n/2$.
  - $\langle d \rangle \geq \frac{n}{8(n-1)} \left\langle k^2 \right\rangle + \frac{1}{2}$

Length in random linear arrangements

- The number of pairs of edges at distance d is $N(d) = n - d$.
- The probability that an edge has length d is [Ferrer-i-Cancho, 2004]

$$p(d) = \frac{N(d)}{\sum_{d=1}^{n-1} N(d)} = \frac{2(n-d)}{n(n-1)} \tag{6}$$

- $E[\langle d \rangle] = E[d] = \frac{n+1}{3}$. Hint: $\sum_{d=1}^{n-1} d^2 = \frac{(n-1)n(2n-1)}{6}$
- $V[d] = \frac{(n+1)(n-2)}{18}$ [Ferrer-i-Cancho, 2013]

Upper bound of $\langle d \rangle$ on non-crossing trees

Outline

- Examples of non-crossing linear arrangements with $\langle d \rangle = n/2$ (star tree and linear tree).
- Prove that $\langle d \rangle = n/2$ is maximum for a non-crossing tree (proof by induction on n). Idea: decomposition of a non-crossing tree into smaller non-crossing trees.

Lower bounds of $\langle d \rangle$ on trees I

The degree method [Petit, 2003]

$$\langle d \rangle = \frac{1}{2(n-1)} \sum_{v=1}^{n} D_v \tag{7}$$

Idea to bound $\langle d \rangle$ below: minimize each D_v (each node v forms a star tree of $n = k_v + 1$ nodes).

If k_v is even

$$D_v \geq \frac{k_v}{2}\left(\frac{k_v}{2} + 1\right) = \frac{k_v^2}{4} + \frac{k_v}{2} \tag{8}$$

If k_v is odd

$$D_v \geq \left(\frac{k_v + 1}{2}\right)^2 = \frac{k_v^2}{4} + \frac{k_v}{2} + \frac{1}{4} \tag{9}$$

Lower bounds of $\langle d \rangle$ on trees II

$$\begin{aligned}
\langle d \rangle &\geq \frac{1}{4(n-1)} \sum_{v=1}^{n} \left(\frac{k_v^2}{2} + k_v \right). & (10) \\
&= \frac{1}{8(n-1)} \sum_{v=1}^{n} k_v^2 + \frac{1}{4(n-1)} \sum_{v=1}^{n} k_v & (11) \\
&= \frac{n}{8(n-1)} \left\langle k^2 \right\rangle + \frac{1}{2}. & (12)
\end{aligned}$$

The importance of star trees: $\langle d \rangle_{min} \leq \langle d \rangle_{min}^{star}$ [Esteban et al., 2016].
More methods to bound $\langle d \rangle$ below [Petit, 2003].

Why is $\langle d \rangle$ below chance in real dependency networks?

A hypothesis on the limited resources of the human brain [Ferrer-i-Cancho, 2004]

- Two linked vertices u and v, such that $\pi(u) < \pi(v)$, the distance $d = \pi(v) - \pi(u)$ can be seen as the time that is needed to keep the open or unresolved dependency in online memory once u has appeared [Morrill, 2000].
- $d = \pi(u) < \pi(v)$ is being minimized, but how exactly?

A family of models to consider:

- minimum linear arrangement problem (sum of dependency lengths)
- minimum bandwidth problem (minimize maximum dependency length)
- ...

The minimum linear arrangement problem [Díaz et al., 2002]

- $u \sim v$ indicates an edge between vertices u and v.
- Find π such that

$$D = \sum_{u \sim v} |\pi(u) - \pi(v)| \tag{13}$$

  is minimum.
- $D = \langle d \rangle / E$. In a tree: $D = \langle d \rangle / (n-1)$.
- Computational complexity:
  - NP-complete for an unconstrained graph [Garey and Johnson, 1979].
  - Polynomial time for a tree.

Minimum linear arrangements of trees

Unconstrained [Petit, 2011]:

- $O(n^3)$ [Goldberg and Klipker, 1976]
- $O(n^{2.2})$ [Shiloach, 1979]
- $O(n^\lambda)$, with $\lambda > \frac{\log 3}{\log 2} = 1.585...$ [Chung, 1984]

Constrained:

- Non-crossing trees: $O(n)$ [Hochberg and Stallmann, 2003].
- Complete k-level 3-ary trees: $O(n)$ [Chung, 1981].
- More examples... [Petit, 2011].

Big question: is a linear time algorithm for unrestricted trees possible?

Experiment

For a given n,

- Produce many random (labelled) trees.
- Arrange the vertices linearly in an arbitrary order and obtain $\langle d \rangle_0$.
- Arrange the vertices linearly solving the minimum linear arrangement problem to obtain $\langle d \rangle_{mla}$.
- What predictions can we make about $\langle d \rangle_0$ and $\langle d \rangle_{mla}$?

An example: Fig. 2 a) of [Ferrer-i-Cancho, 2006].

- Power-laws? $\rightarrow$ Model selection.
- Producing uniformly distributed random trees: the Aldous-Brother algorithm [Aldous, 1990, Broder, 1989].
- What is the mathematical form of $\langle d \rangle_{mla}$? Theoretical and **empirical** approach.

Interest of crossings

- Computational efficiency (m.l.a. without crossings in linear time [Hochberg and Stallmann, 2003]).
- Theoretical linguistics, computational linguistics and cognitive science.
  - Projectivity = planarity + uncovered root (context-freeness) [Mel'čuk, 1988]
  - Mild context-sensitivity [Joshi, 1985]
- ...

The maximum number of crossings I

- Q: the set of pairs of edges that may potentially cross.
- C: the number of edge crossings, $C \leq |Q|$

$$|Q| = \frac{1}{4} \sum_{u=1}^{n} \sum_{v=1}^{n} a_{uv} C_{pairs}(u, v) \tag{14}$$

The number of crossings in which the edge $u \sim v$ is involved cannot exceed

$$C_{pairs}(u, v) = n - k_u - k_v, \tag{15}$$

being k_v the degree of vertex v.

C defines the number of pairs of edges that can cross.

Notice the 1/4 factor of Eq. 14.

The maximum number of crossings II

$$|Q| = \frac{n}{2}\left(n - 1 - \langle k^2 \rangle\right) \tag{16}$$

- Given n, $|Q|$ is determined by $\langle k^2 \rangle$.
- $|Q| \geq 0$ yields $\langle k^2 \rangle \leq n - 1$. What are the trees for which $\langle k^2 \rangle = n - 1$?
- What are the trees minimizing $\langle k^2 \rangle$?

The expected number of crossings I

$p_c(u, v; s, t)$ is the probability that the edges $u \sim v$ and $s \sim t$ cross.

- $p_c(u, v; s, t) = 0$ if $u \sim v$ and $s \sim t$ share at least one vertex.
- $p_c(u, v; s, t) = 1/3$ otherwise. Outline:
  - Generate four different random numbers from 1 to n.
  - Sort them increasingly.
  - Choose the position of the vertices of one the edges.
  - Then

$$p_c(u, v; s, t) = \frac{2}{\binom{4}{2}} = \frac{1}{3} \tag{17}$$

The expected number of crossings II

Decomposition of C as a sum of indicator variables

$$C = \frac{1}{4} \sum_{u=1}^{n} \sum_{v=1}^{n} a_{uv} C(u, v) \tag{18}$$

and

$$C(u, v) = \frac{1}{2} \sum_{\substack{s = 1 \\ s \neq u, v}}^{n} \sum_{\substack{t = 1 \\ t \neq u, v}}^{n} a_{st} C(u, v; s, t) \tag{19}$$

with $C(u, v; s, t) \in \{0, 1\}$.

The expected number of crossings III

The expectation of the sum is the sum of expectations

$$E[C] = \frac{1}{4} \sum_{u=1}^{n} \sum_{v=1}^{n} a_{uv} E[C(u,v)] \tag{20}$$

and

$$E[C(u,v)] = \frac{1}{2} \sum_{\substack{s=1 \\ s \neq u,v}}^{n} \sum_{\substack{t=1 \\ t \neq u,v}}^{n} a_{st} E[C(u,v;s,t)] \tag{21}$$

with $E[C(u,v;s,t)] = p_c(u,v;s,t)$.

The expected number of crossings IV

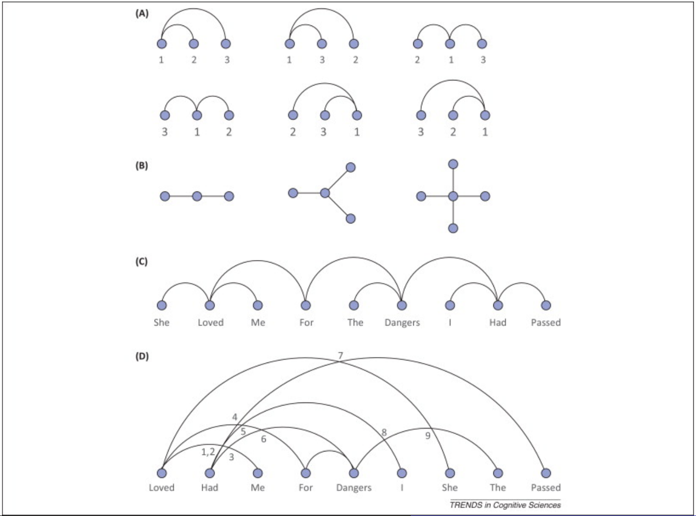
Thus,

$$E[C] = |Q| p_c(u, v; s, t) \tag{22}$$

$$= \frac{|Q|}{3} \tag{23}$$

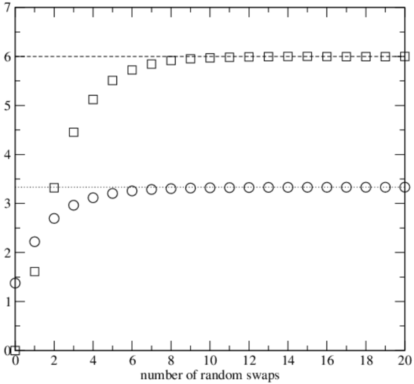
$$= \frac{n}{6}\left(n - 1 - \left\langle k^2 \right\rangle\right) \tag{24}$$

Does the theory work? I



Does the theory work? II

Progressive randomization of the vertex sequence [Ferrer-i-Cancho, 2017]



- Initial example of an English sentence ($n = 9$) starting at
  - $\langle d \rangle = 11/8 = 1.375$
  - $C = 0$.
- Circles: $\langle d \rangle \to \frac{n+1}{3} = 10/3$.
- Squares: $C \to \frac{n}{6}\left(n - 1 - \langle k^2 \rangle\right) = 6$.

Positive correlation between $\langle d \rangle$ and C!

Crossings in uniformly random trees I

$$E[C] = \frac{n}{6}\left(n - 1 - \left\langle k^2 \right\rangle\right) \quad (25)$$

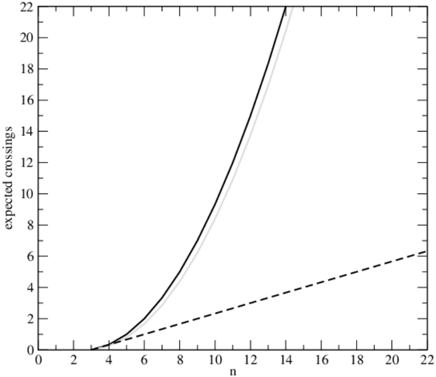
The degree variance for uniformly random labelled trees [Moon, 1970, Noy, 1998]

$$V[k] = \left\langle k^2 \right\rangle - \left\langle k \right\rangle^2 = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \quad (26)$$

Applying $\langle k \rangle = 2 - 2/n$ yields

$$\left\langle k^2 \right\rangle = \frac{n-1}{n}\left(5 - \frac{6}{n}\right) \quad (27)$$

Crossings in uniformly random trees II



There must be a hidden constraint for the scarcity of crossings in real sentences [Ferrer-i-Cancho, 2016]

- Linear trees.
- Uniformly random labelled trees.
- (quasi-star trees)
- Star trees?

An more precise null hypothesis predicts the actual number of crossings with a relative error that is not greater than about 5% (on average)! [Ferrer-i-Cancho, 2014, Gómez-Rodríguez and Ferrer-i-Cancho, 2016].

To conclude

- Real data suggest that $\langle d \rangle$ is been minimized in real trees.
- The small values of C in real dependency trees might be a side-effect of the minimization of $\langle d \rangle$. Figs. 2 c) and d) of [Ferrer-i-Cancho, 2006]
- It is not known how this optimization actually works (but sentence production is not a batch process [Christiansen and Chater, 2016]).
- A mathematical description of $\langle d \rangle$ and C as a function of n in real dependency trees or optimized (mla) trees is not forthcoming.

Aldous, D. (1990).
The random walk construction of uniform spanning trees and uniform labelled trees.
SIAM J. Disc. Math., 3:450–465.

Broder, A. (1989).
Generating random spanning trees.
In *Symp. Foundations of Computer Sci., IEEE*, pages 442–447, New York.

Christiansen, M. H. and Chater, N. (2016).
The now-or-never bottleneck: a fundamental constraint on language.
Behavioral & Brain Sciences, 39:e62.

Chung, F. R. K. (1981).
Some problems and results in labelings of graphs.

In Chartrand, G., editor, *The Theory and Applications of Graphs*, page 255264. John Wiley and Sons, New York.

Chung, F. R. K. (1984).
On optimal linear arrangements of trees.
Comp. & Maths. with Appls., 10(1):43–60.

Díaz, J., Petit, J., and Serna, M. (2002).
A survey of graph layout problems.
ACM Computing Surveys, 34:313–356.

Esteban, J. L., Ferrer-i-Cancho, R., and Gómez-Rodríguez, C. (2016).
The scaling of the minimum sum of edge lengths in uniformly random trees.
Journal of Statistical Mechanics, page 063401.

Ferrer-i-Cancho, R. (2004).

Euclidean distance between syntactically linked words.
Physical Review E, 70:056135.

Ferrer-i-Cancho, R. (2006).
Why do syntactic links not cross?
Europhysics Letters, 76(6):1228–1235.

Ferrer-i-Cancho, R. (2013).
Hubiness, length, crossings and their relationships in dependency trees.
Glottometrics, 25:1–21.

Ferrer-i-Cancho, R. (2014).
A stronger null hypothesis for crossing dependencies.
Europhysics Letters, 108:58003.

Ferrer-i-Cancho, R. (2016).
Non-crossing dependencies: least effort, not grammar.

In Mehler, A., Lücking, A., Banisch, S., Blanchard, P., and Job, B., editors, *Towards a theoretical framework for analyzing complex linguistic networks*, pages 203–234. Springer, Berlin.

Ferrer-i-Cancho, R. (2017).
Random crossings in dependency trees.
Glottometrics, 37:1–12.

Garey, M. R. and Johnson, D. S. (1979).
Computers and intractability: a guide to the theory of NP-completeness.
W. M. Freeman, San Francisco.

Goldberg, M. K. and Klipker, I. A. (1976).
Minimal placing of trees on a line.
Technical report, Physico-Technical Institute of Low Temperatures. Academy of Sciences of Ukranian SSR, USSR. in Russian.

Gómez-Rodríguez, C. and Ferrer-i-Cancho, R. (2016).
The scarcity of crossing dependencies: a direct outcome of a specific constraint?
http://arxiv.org/abs/1601.03210.

Hochberg, R. A. and Stallmann, M. F. (2003).
Optimal one-page tree embeddings in linear time.
Information Processing Letters, 87:59–66.

Joshi, A. K. (1985).
Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions?
In *Natural Language Parsing*, page 206250. Cambridge University Press.

Mel'čuk, I. (1988).
Dependency syntax: theory and practice.

State of New York University Press, Albany.

Moon, J. (1970).
Counting labelled trees.
In *Canadian Math. Cong.*

Morrill, G. (2000).
Incremental processing and acceptability.
Computational Linguistics, 25(3):319–338.

Noy, M. (1998).
Enumeration of noncrossing trees on a circle.
Discrete Mathematics, 180:301–313.

Petit, J. (2003).
Experiments on the minimum linear arrangement problem.
J. Exp. Algorithmics, 8.

Petit, J. (2011).

Addenda to the survey of layout problems.
Bulletin of the European Association for Theoretical Computer Science, 105:177–201.

Shiloach, Y. (1979).
A minimum linear arrangement algorithm for undirected trees.
SIAM J. Comput., 8(1):15–32.