Extraction of information in large graphs
Automatic search for synonyms

Pierre Senellart,
under the direction of Prof. Vincent Blondel

June 5th 2001 - August 3rd 2001

The dictionary graph

Computation (n.) The act or process of computing; calculation; reckoning.

Computation (n.) The result of computation; the amount computed.

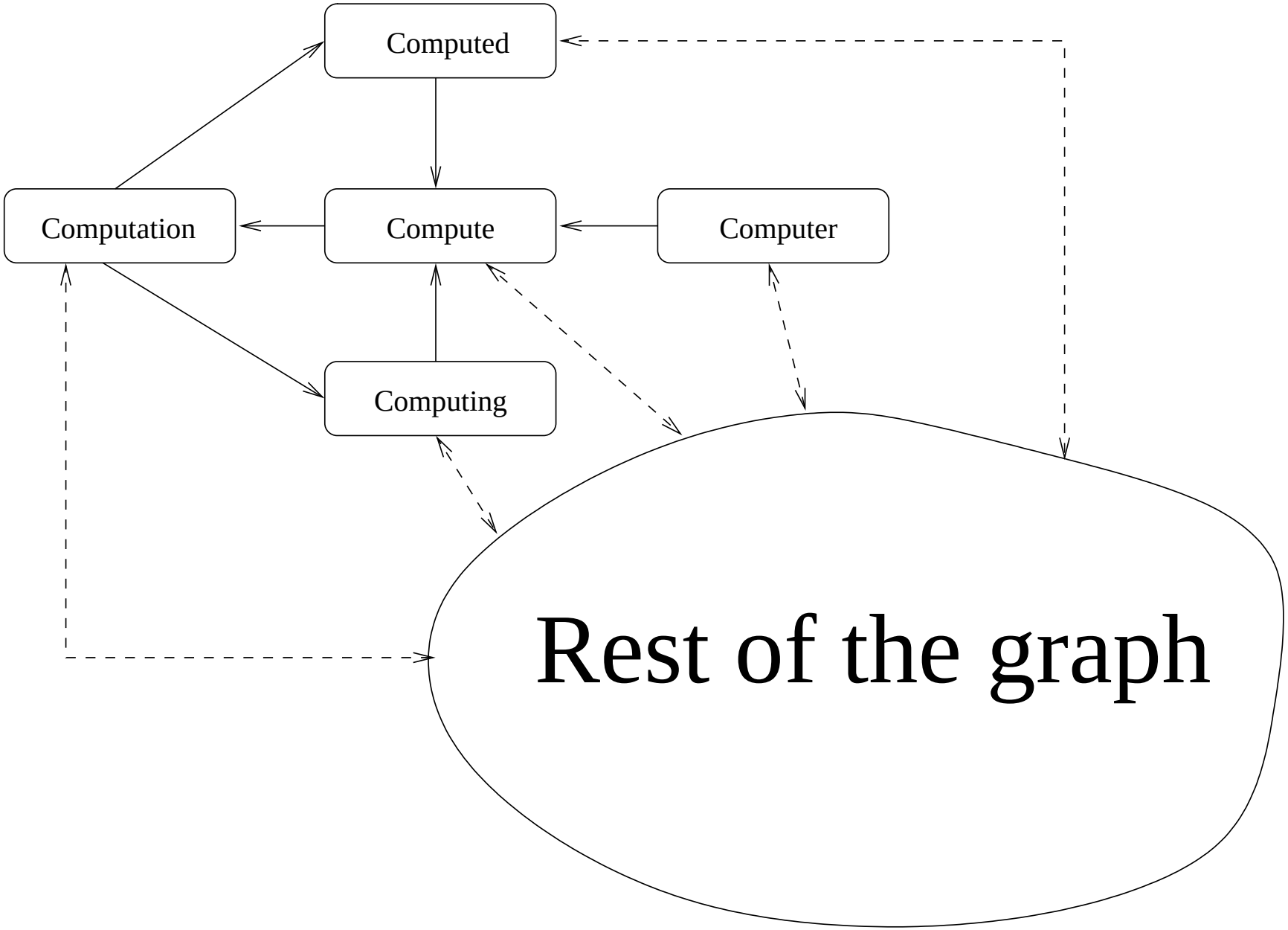
Computed (imp. & p. p.) of Compute

Computing (p. pr. & vb. n.) of Compute

Compute (v.t.) To determine calculation; to reckon; to count.

Compute (n.) Computation.

Computer (n.) One who computes.


Computed
Computation
Compute
Computer
Computing
Rest of the graph

Extraction of the graph

- Multiwords (e.g. All Saints', Surinam toad)
- Prefixes and suffixes (e.g un-, -ous)
- Different meanings of a word
- Derived forms (e.g. daisies, sought)
- Accentuated characters (e.g. proven/al, cr/che)
- Misspelled words

$112,169$ vertices - $1,398,424$ arcs.

Lexical units

$13,396$ lexical units not defined in the dictionary:

- Numbers (e.g. 14159265, 14th)
- Mathematical and chemical symbols (e.g. x3, fe3o4)
- Proper nouns (e.g. California, Aaron)
- Misspelled words (e.g. aligator, abudance)
- Undefined words (e.g. snakelike, unwound)
- Abbreviations (e.g. adj, etc)

Connectivity

185 different connected components:

- 1 111,982-vertex component
- 3 2-vertex components

anguineal⟶anguineous⟶snakelike
indissolvableness⟶indissolubleness⟶indissolubility

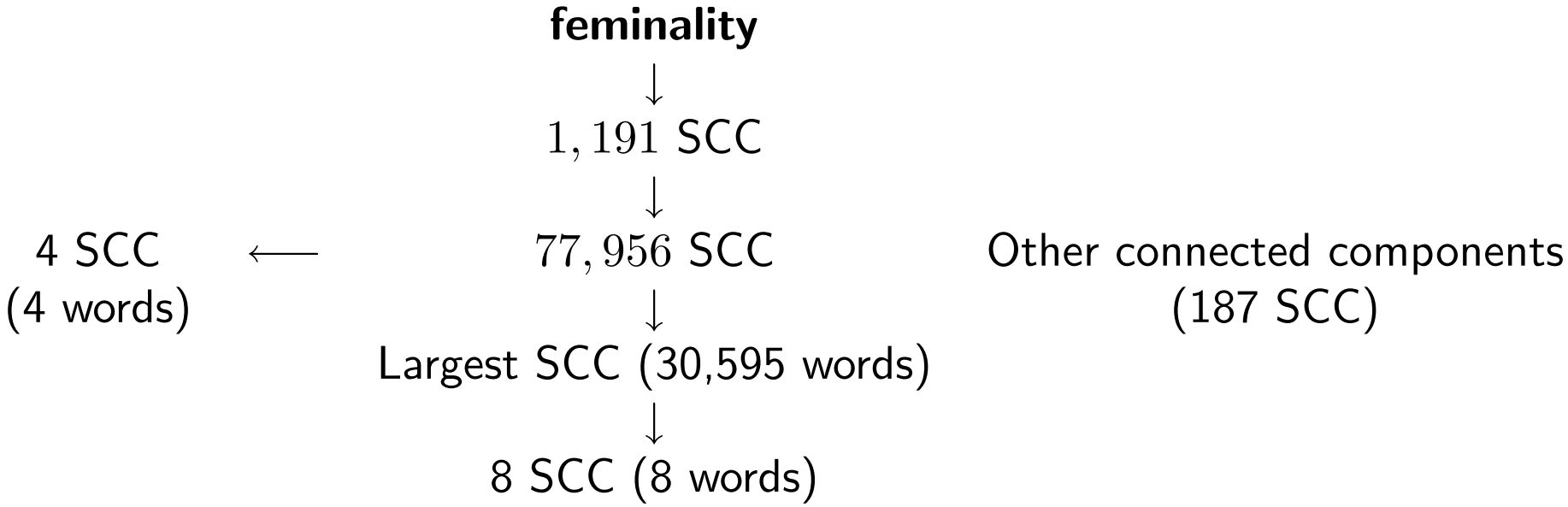
- 181 1-vertex components

Strong connectivity

$79,348$ strongly connected components.

Number of vertices	Number of components
$30,595$	1
10	1
7	3
6	13
5	21
4	50
3	222
2	$1,457$
1	$77,580$

Distribution of the size of strongly connected components



Graph resulting of the contraction of each SCC in one single vertex

10-vertex component: bezpopovtsy, dukhobors, dukhobortsy, judaizers, molokane, molokany, popovtsy, raskolnik, raskolniki, skoptsy.

A core of the language

Definition. *A* core subgraph *of a graph G is a subgraph of G with the two following properties:*

1. *It contains at least one vertex from every directed cycle in G.*

2. *Every path in the graph may be prolongated in a path containing a vertex of the subgraph.*

If you know the meaning of all the words in a core subgraph of the dictionary graph, you may learn the meaning of all words in the dictionary.

The largest SCC (plus the other connected components and 12 words) is a core subgraph of the dictionary.

The independence degree

Definition. *The* independence degree *of a graph G is the minimum number of vertices of a core subgraph of G.*

Theorem. *The computation of the independence degree of a graph is a $NP - complete$ problem.*

Upper bound : $30,905 + 187 + 12 = 30,794$.

Good approximation algorithm?

A small world

Definition. *A graph is a small world graph if it has the following properties:*

1. *It is undirected, unweighted, sparse and connected.*

2. *The mean minimal length of a path between any two vertices (which is called the* characteristic path length*) L is close to that of a random graph with same n and k.*

3. *The mean over all vertices of the ratio of the number of edges in the neighborhood graph by the number of possible edges in the same subgraph (which is called the* clustering coefficient*) γ is much greater than that of a random graph of same n and k.*

The Web, power distribution graphs, the Kevin Bacon graph are well-known examples of small worlds.

The underlying undirected graph of the largest connected component of the dictionary graph *is* a small world:

1. Obvious.

2. $L \approx 2,40 \ \sim \ 3.61 \approx L_{\text{random}}$

3. $\gamma \approx 0.45 \ \gg \ 2.19 \ 10^{-4} \approx \gamma_{\text{random}}$

Yet it does not fit the models of small worlds graphs proposed by Duncan J. Watts.

Necessity of a model of directed small worlds?

Degree distributions

A Zipfian dustribution: the probability that a node has indegree or outdegree i is proportional to $1/i^{\alpha}$ for some α.

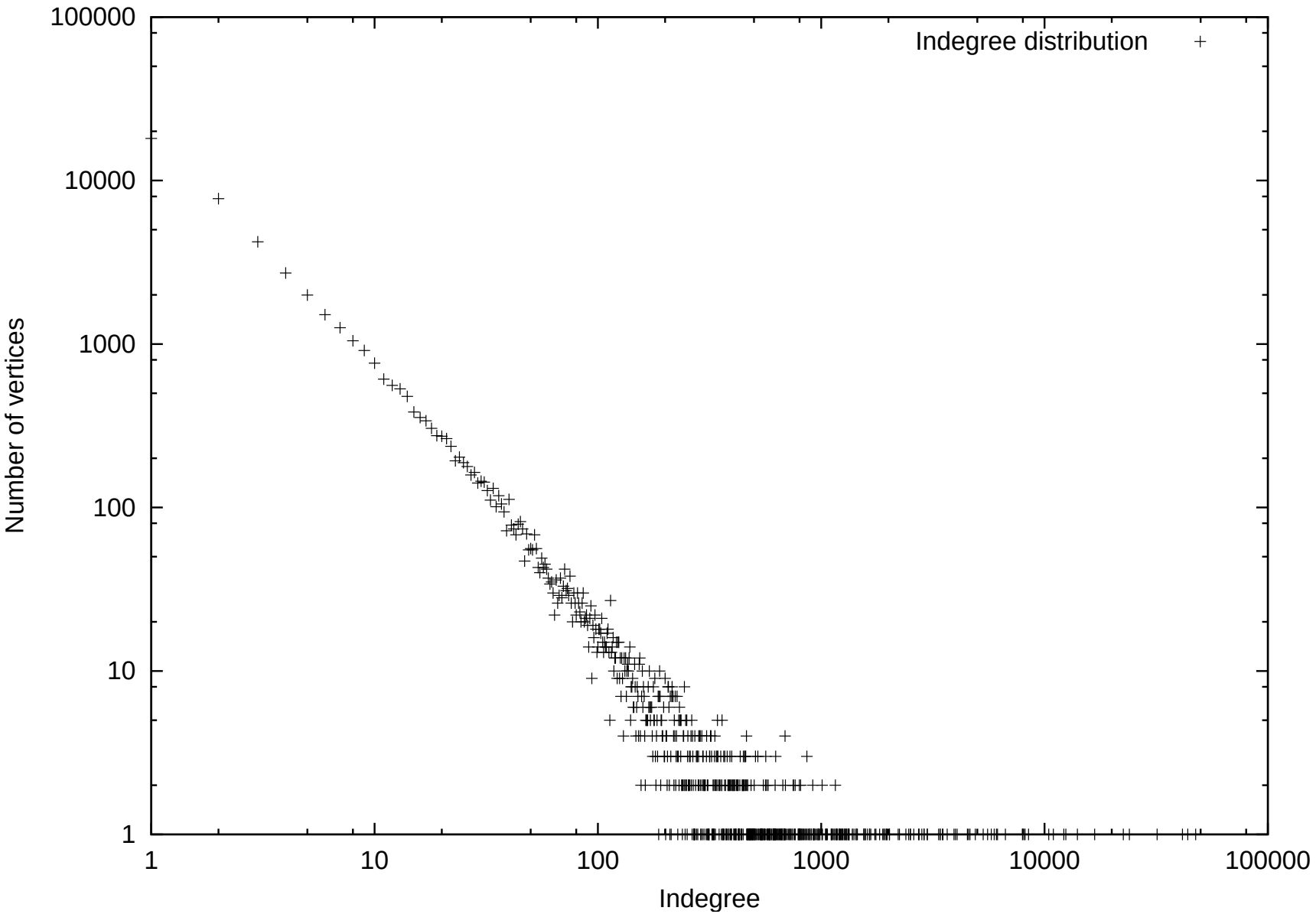
Indegree : $\alpha \approx 1.6$. Outdegree : $\alpha \approx 3.1$

Concerning the outdegree distribution:

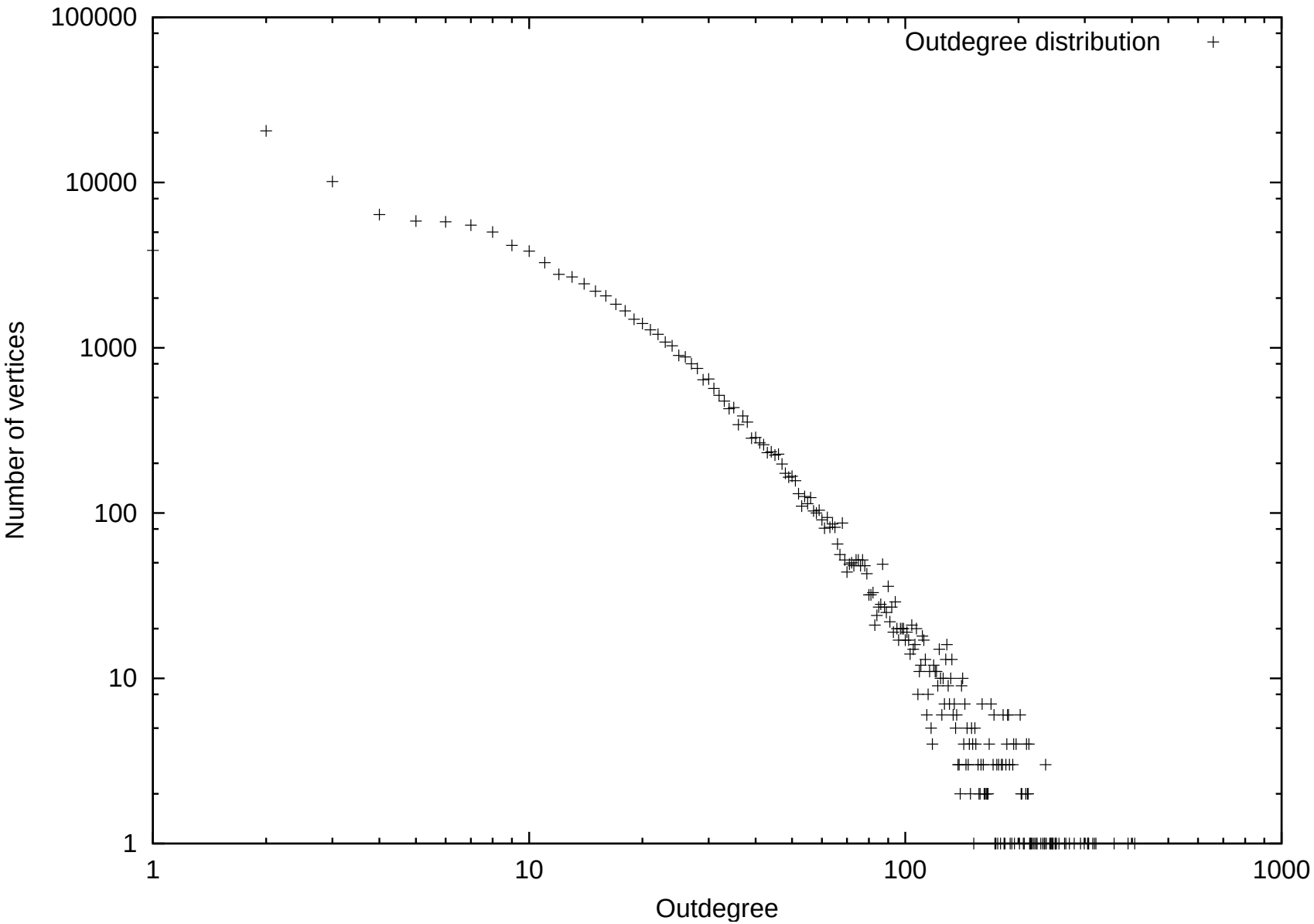
1. It is bounded by a rather small ammount.

2. The plot is not linear in the range of small outdegrees.

Same kind of distributions as for the Web.



Indegree distribution



Outdegree distribution

Looking for near-synonyms

Definition. *The* neighborhood graph *of a node i in a directed graph G is the subgraph consisting of i, all parents of i and all children of i.*

i is some word we want a synonym of.

A will be the adjacency matrix of the neighborhood graph of i in the dictionary graph.

n is the order of A.

The vectors method

For each $1 \leq j \leq n, j \neq i$, compute:

$$\|(A_{i,.} - A_{j,.})\| + \|(A_{.,i} - A_{.,j})^T\|$$

(where $\| \ \|$ is some vector norm, $A_{i,.}$ and $A_{.,i}$ are respectively the ith line and the ith column of A).

For instance, if we choose the Euclidean norm, we compute:

$$\left(\sum_{k=1}^{n}(A_{i,k} - A_{j,k})^2\right)^{\frac{1}{2}} + \left(\sum_{k=1}^{n}(A_{k,i} - A_{k,j})^2\right)^{\frac{1}{2}}$$

The lower this value is, the best j is a synonym of i.

Kleinberg's algorithm

$$Hub \longrightarrow Authority$$

A *mutually reinforcing relationship*: good hubs are pages that point to good authorities and good authorities are pages pointed by good hubs.

The principal eigenvectors of $A^T A$ and AA^T give respectively the *authority weights* and *hub weights* of the vertices of the graph.

An extension of Kleinberg's algorithm

Let $M(m,m)$ and $N(n,n)$ be the transition matrices of two oriented graphs.

Let $C = M \otimes N + M^T \otimes N^T$ where $\otimes$ is the Kronecker tensorial product.

We assume that the greatest eigenvalue of C is strictly greater than the absolute value of all other eigenvalues.

Then, the normalized principal eigenvector X of C gives the "similarity" between a vertex of M and a vertex of N: $X_{i \times n+j}$ characterizes the similarity between vertex i of M and vertex j of N.

In particular, if $M = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, the result is that of Kleinberg's algorithm.

Application to the search for synonyms

$$1 \longrightarrow 2 \longrightarrow 3$$

We are looking for vertices "like 2" in the neighborhood graph of i.

Let $C = M \otimes A + M^T \otimes A^T$ where $M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$.

The principal eigenvector of C gives the similarity between a node in G and a node in the graph $1 \longrightarrow 2 \longrightarrow 3$.

We just select the subvector corresponding to the vertex 2 in order to have synonymy weights.

ArcRank

PageRank (Google): stationary distribution of weights over vertices corresponding to the principal eigenvector of the adjacency matrix.

ArcRank:

$$r_{s,t} = \frac{p_s/|a_s|}{p_t}$$

$|a_s|$ is the outdegree of s.

p_t is the pagerank of t.

The best synonyms of i are the other extremity of the best-ranked arcs arriving to or leaving from i.

Disappear

	Vectors	Kleinberg	ArcRank	Wordnet	Microsoft Word
1	**vanish**	**vanish**	**epidemic**	**vanish**	**vanish**
2	**wear**	**pass**	**disappearing**	**go away**	**cease to exist**
3	**die**	**die**	**port**	**end**	**fade away**
4	**sail**	**wear**	**dissipate**	**finish**	**die out**
5	**faint**	**faint**	**cease**	**terminate**	**go**
6	**light**	**fade**	**eat**	**cease**	**evaporate**
7	**port**	**sail**	**gradually**		**wane**
8	**absorb**	**light**	**instrumental**		**expire**
9	**appear**	**dissipate**	**darkness**		**withdraw**
10	**cease**	**cease**	**efface**		**pass away**

Table 1: Near-synonyms for **disappear**

Parallelogram

	Vectors	Kleinberg	ArcRank	Wordnet	Microsoft Word
1	**square**	**square**	**quadrilateral**	**quadrilateral**	**diamond**
2	**parallel**	**rhomb**	**gnomon**	**quadrangle**	**lozenge**
3	**rhomb**	**parallel**	**right-lined**	**tetragon**	**rhomb**
4	**prism**	**figure**	**rectangle**		
5	**figure**	**prism**	**consequently**		
6	**equal**	**equal**	**parallelopiped**		
7	**quadrilateral**	**opposite**	**parallel**		
8	**opposite**	**angles**	**cylinder**		
9	**altitude**	**quadrilateral**	**popular**		
10	**parallelopiped**	**rectangle**	**prism**		

Table 2: Near-synonyms for **parallelogram**

Sugar

	Vectors	Kleinberg	ArcRank	Wordnet	Microsoft Word
1	**juice**	**cane**	**granulation**	**sweetening**	**darling**
2	**starch**	**starch**	**shrub**	**sweetener**	**baby**
2	**cane**	**sucrose**	**sucrose**	**carbohydrate**	**honey**
4	**milk**	**milk**	**preserve**	**saccharide**	**dear**
5	**molasses**	**sweet**	**honeyed**	**organic compound**	**love**
6	**sucrose**	**dextrose**	**property**	**saccarify**	**dearest**
7	**wax**	**molasses**	**sorghum**	**sweeten**	**beloved**
8	**root**	**juice**	**grocer**	**dulcify**	**precious**
9	**crystalline**	**glucose**	**acetate**	**edulcorate**	**pet**
10	**confection**	**lactose**	**saccharine**	**dulcorate**	**babe**

Table 3: Near-synonyms for **sugar**

Science

	Vectors	Kleinberg	ArcRank	Wordnet	Microsoft Word
1	**art**	**art**	**formulate**	**knowledge domain**	**discipline**
2	**branch**	**branch**	**arithmetic**	**knowledge base**	**knowledge**
3	**nature**	**law**	**systematize**	**discipline**	**skill**
4	**law**	**study**	**scientific**	**subject**	**art**
5	**knowledge**	**practice**	**knowledge**	**subject area**	
6	**principle**	**natural**	**geometry**	**subject field**	
7	**life**	**knowledge**	**philosophical**	**field**	
8	**natural**	**learning**	**learning**	**field of study**	
9	**electricity**	**theory**	**expertness**	**ability**	
10	**biology**	**principle**	**mathematics**	**power**	

Table 4: Near-synonyms for **science**

Perspectives

- Extension of the subgraph
- Other dictionaries, other languages
- Other applications of the extension of Kleinberg's algorithm
- A model of small world directed graphs
- Invariants for languages