AUTHORSHIP RECOGNITION USING THE DYNAMICS OF CO-OCCURRENCE NETWORKS

Camilo Akimushkin Valencia
camilo.akimushkin@gmail.com

Instituto de Física de São Carlos
Universidade de São Paulo

11 November 2016

Authorship recognition

Lexical features (frequency)

Words, n-grams, functional words, types of words, discourse-connecting expressions, slang, contractions, dialects, orthography mistakes, proper names, semantic features (polysemy).

Text format

Lengths of lines, words and phrases, formatting (white spaces), capitalization, non-alphanumeric characters, beginnings and ends of texts.

Character-level features

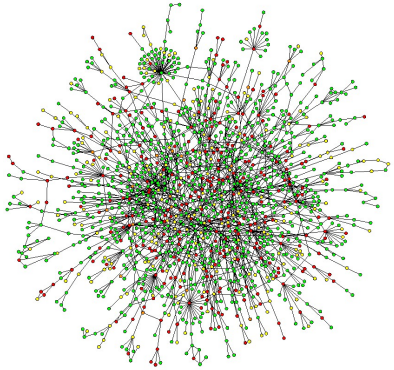
Character n-grams, frequent suffixes, punctuation.

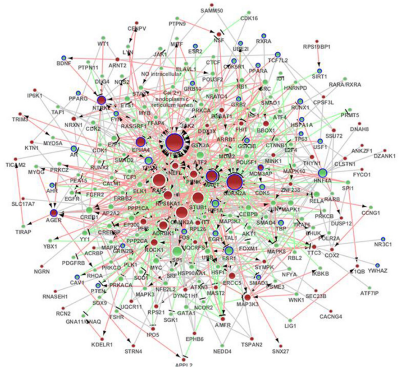
Other

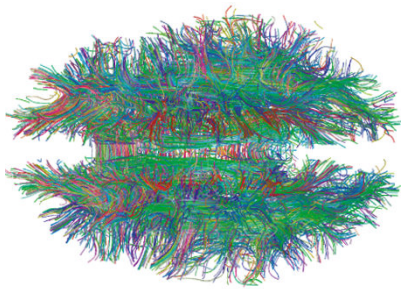
Syntactic features: n-grams syntactic function, kinds of phrases, perplexity, morphological complexity.

Complex networks

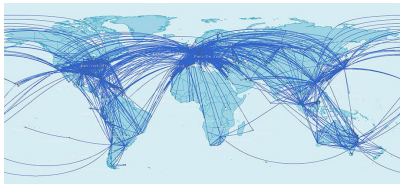
Natural

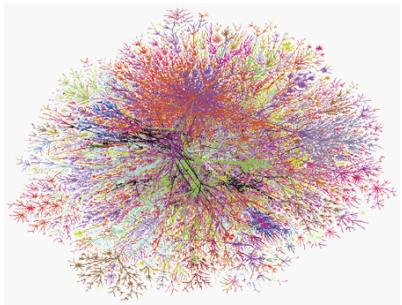


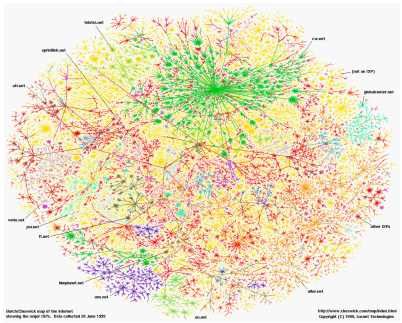




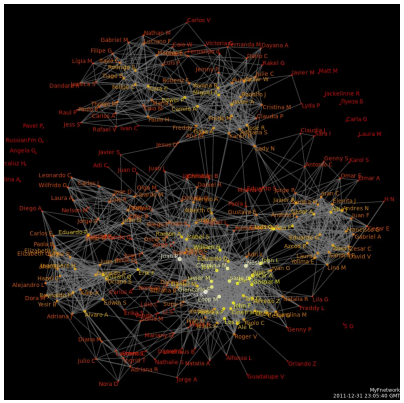
Artificial

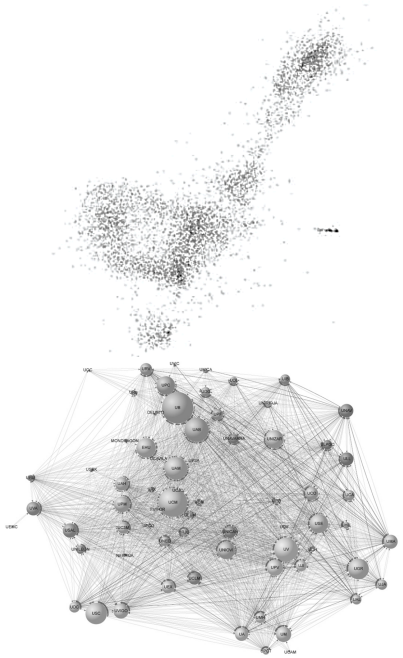




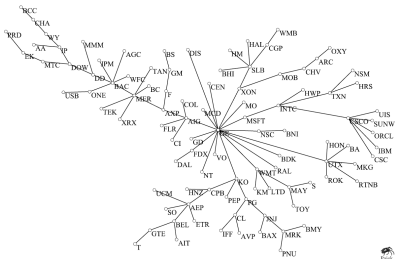


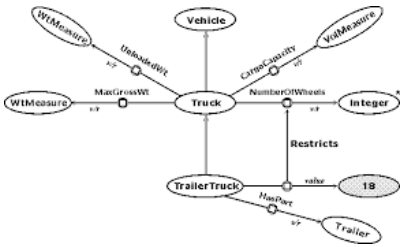
Social

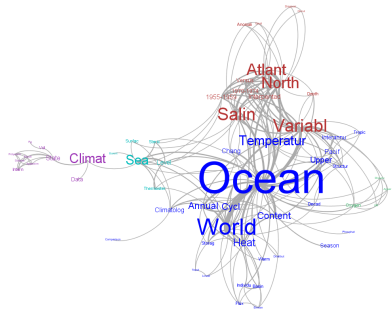




Other



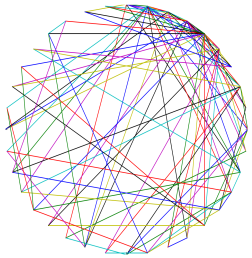


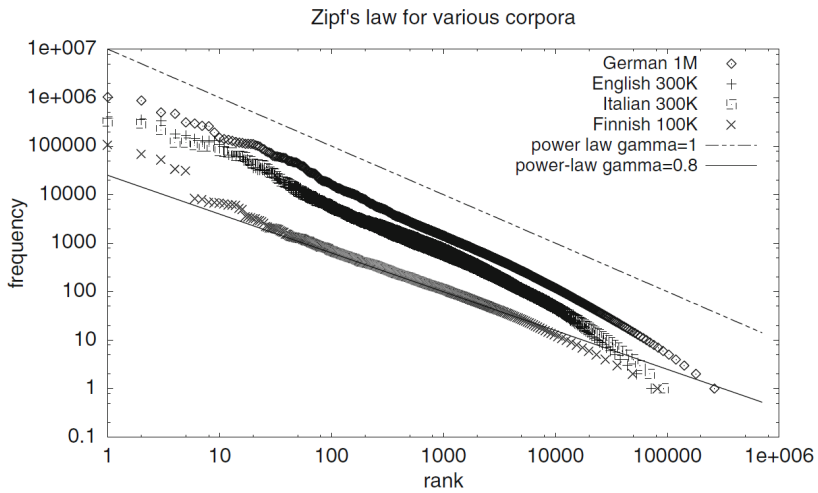


Complexity of language

Zipf's law (1939)

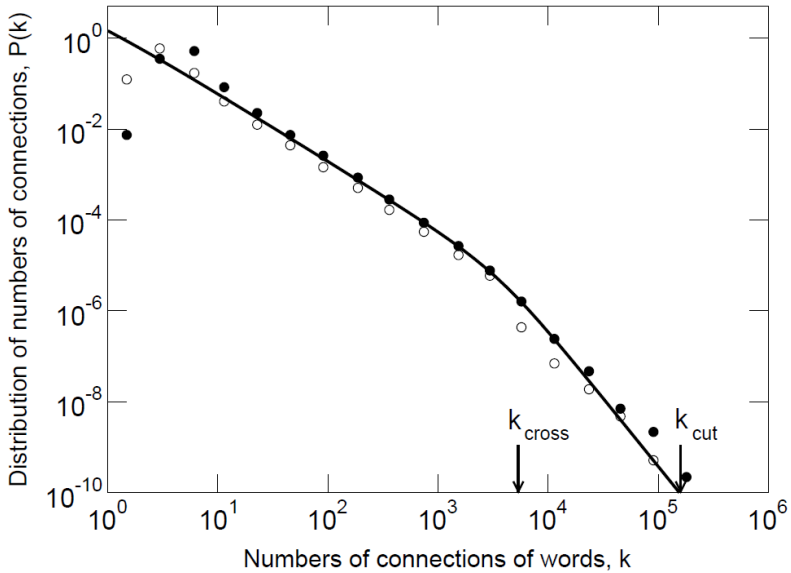
$$f_j = A r_j^{-q}, \quad q = 1$$





Biemann, Quasthoff. Networks generated from natural language text. In: *Dynamics on and of Complex Networks.* **2009.**

$$P(k) \sim k^{-\gamma}, \quad \gamma = 1 + q^{-1}$$



Dorogovtsev, Mendes. Language as an Evolving Word Web. *P. Roy. Soc. Lond. B. Bio.* **2001.**

Word co-occurrence networks

Construction

It was the best of times,

it was the worst of times,

it was the age of wisdom,

it was the age of foolishness...

A Tale of Two Cities - Charles Dickens

Word co-occurrence networks

Construction

best times

worst times

age wisdom

age foolishness

Word co-occurrence networks

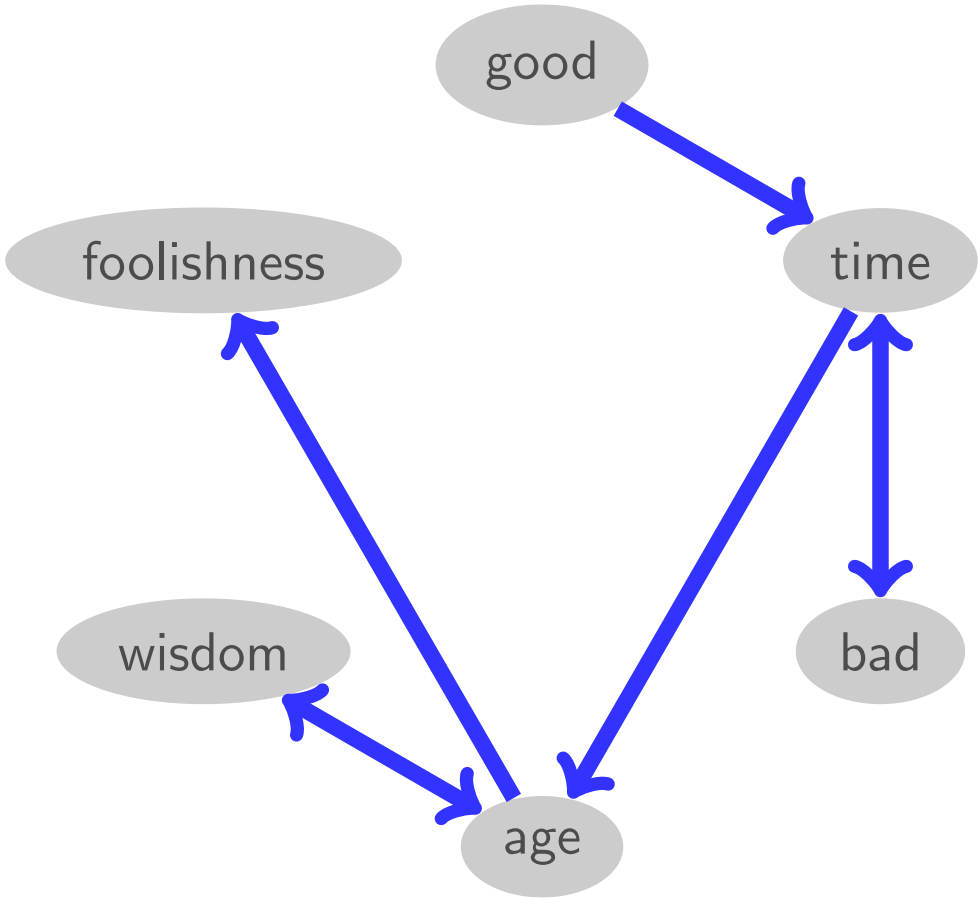
Construction

best times

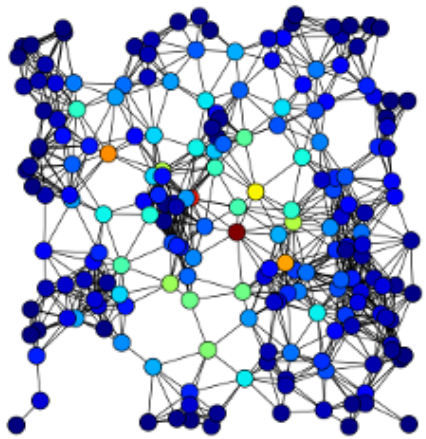
worst times

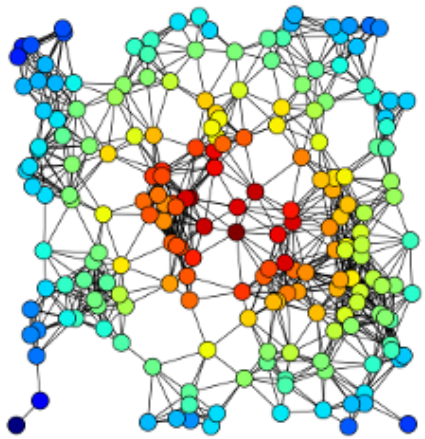
age wisdom

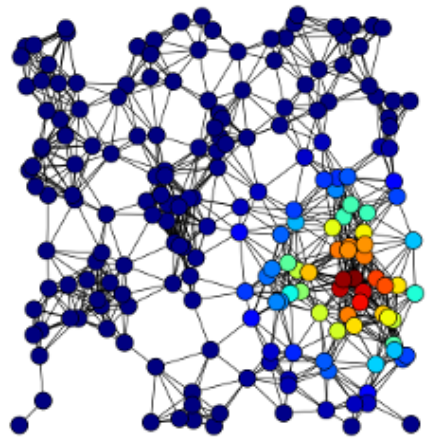
age foolishness

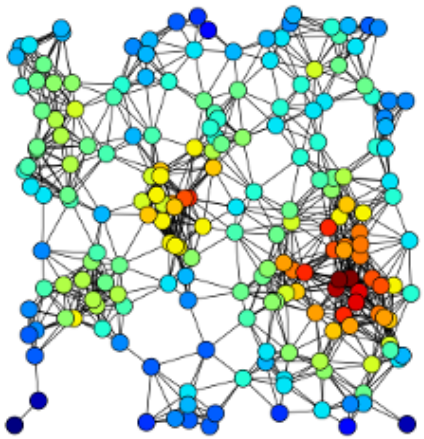


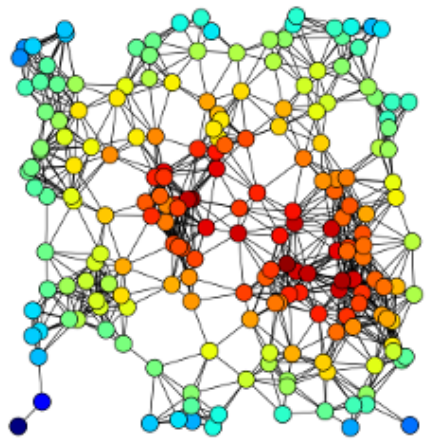
Network metrics

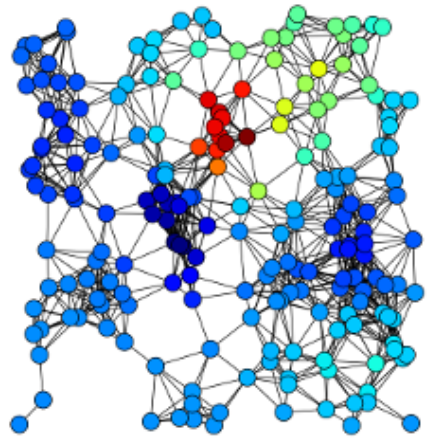

A


B


C


D


E


F

Network metrics

1. Clustering: $c_i = \frac{e_i}{k_i(k_i-1)}$
2. Diameter: $D = \max\{D_{ij}\}$
3. Radius: $r = \min\{D_{ij}\}$
4. Cliques: Number of complete subgraphs
5. Load centrality: Betweenness centrality with loads on edges
6. Transitivity: $T = 3\frac{triangles}{triads}$
7. Betweenness centrality: $B_i = \Sigma_{s\neq i\neq t}\frac{g^i_{st}}{g_{st}}$
8. Shortest path: $l_{ij} = \left[A^n\right]_{ij}$
9. Connectivity: $k_i = \left[A^2\right]_{ii}$
10. Intermittency: $I_i = \mathrm{var}(\Delta)/\bar{\Delta}$
11. Number of nodes: N
12. Number of edges: E

A: Adjacency matrix; $g_{st} = \Sigma l_{st}$; Δ_i: Distance between two appearances of a word.

Dynamics of networks for authorship recognition

Authorship of books

- Few books per author
- Depends on style
- Small networks
- Uneven networks

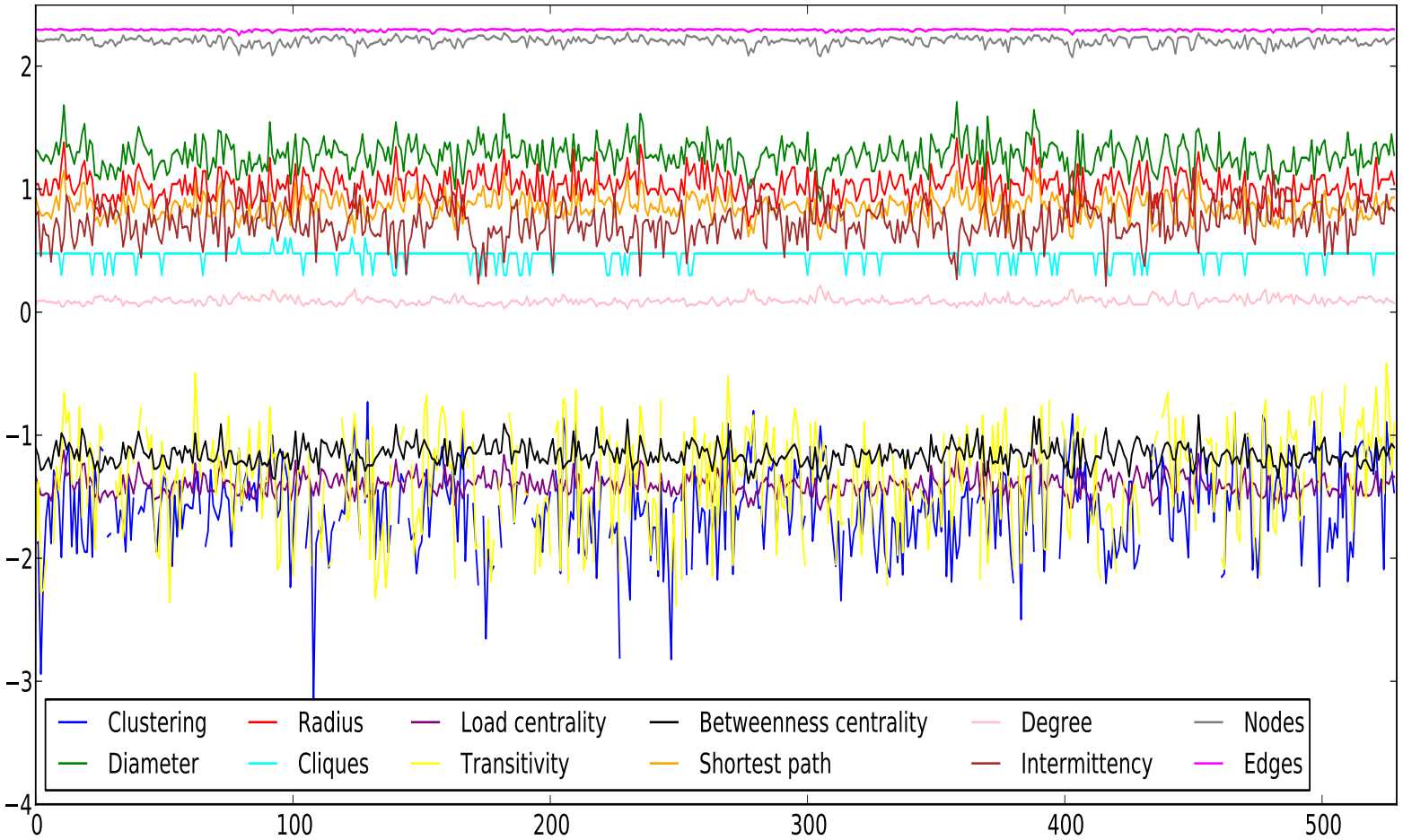
Dynamics of networks for authorship recognition

Authorship of books

- Few books per author
- Depends on style
- Small networks
- Uneven networks

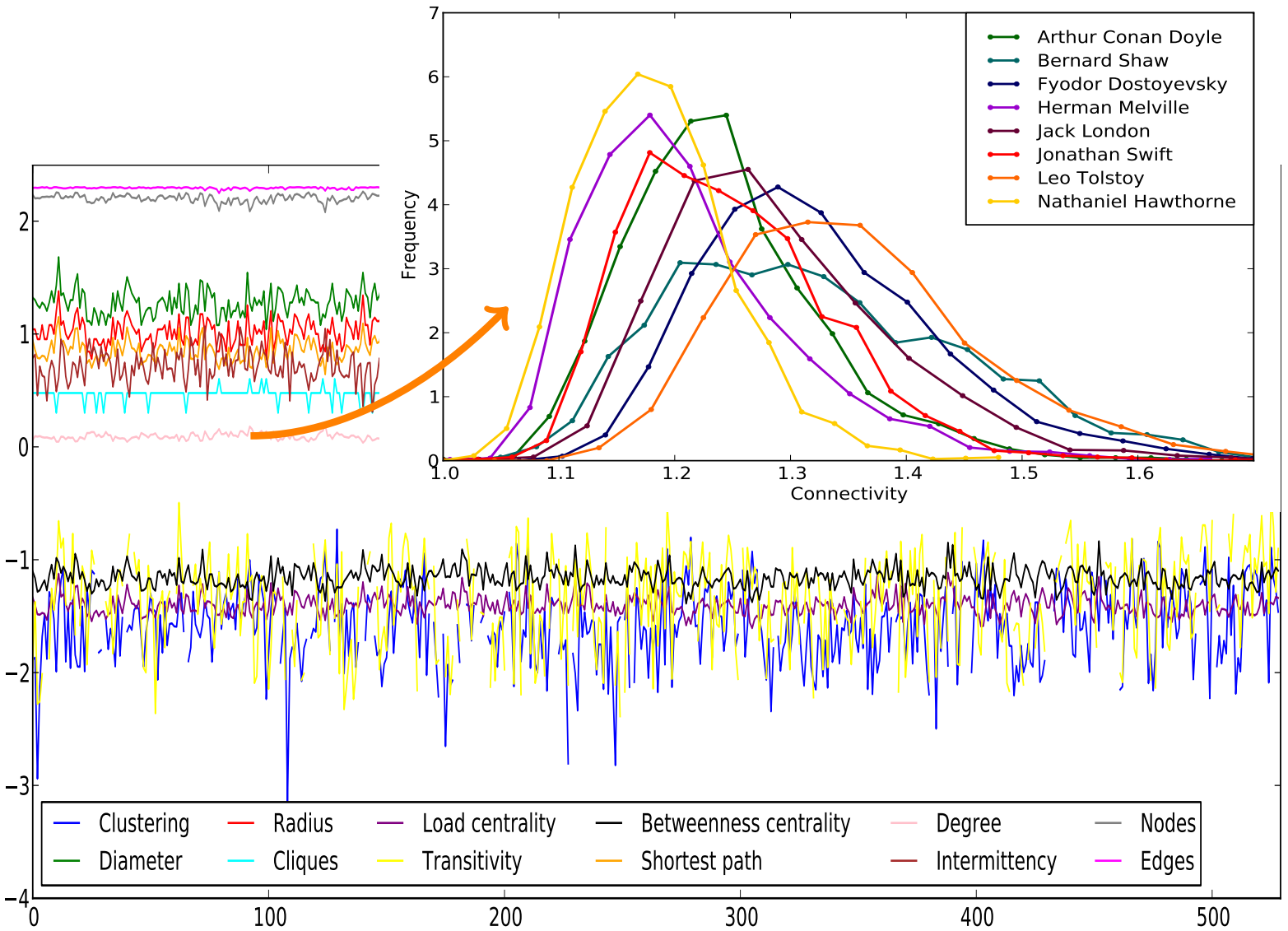
"Dynamics" OF the network

Time series



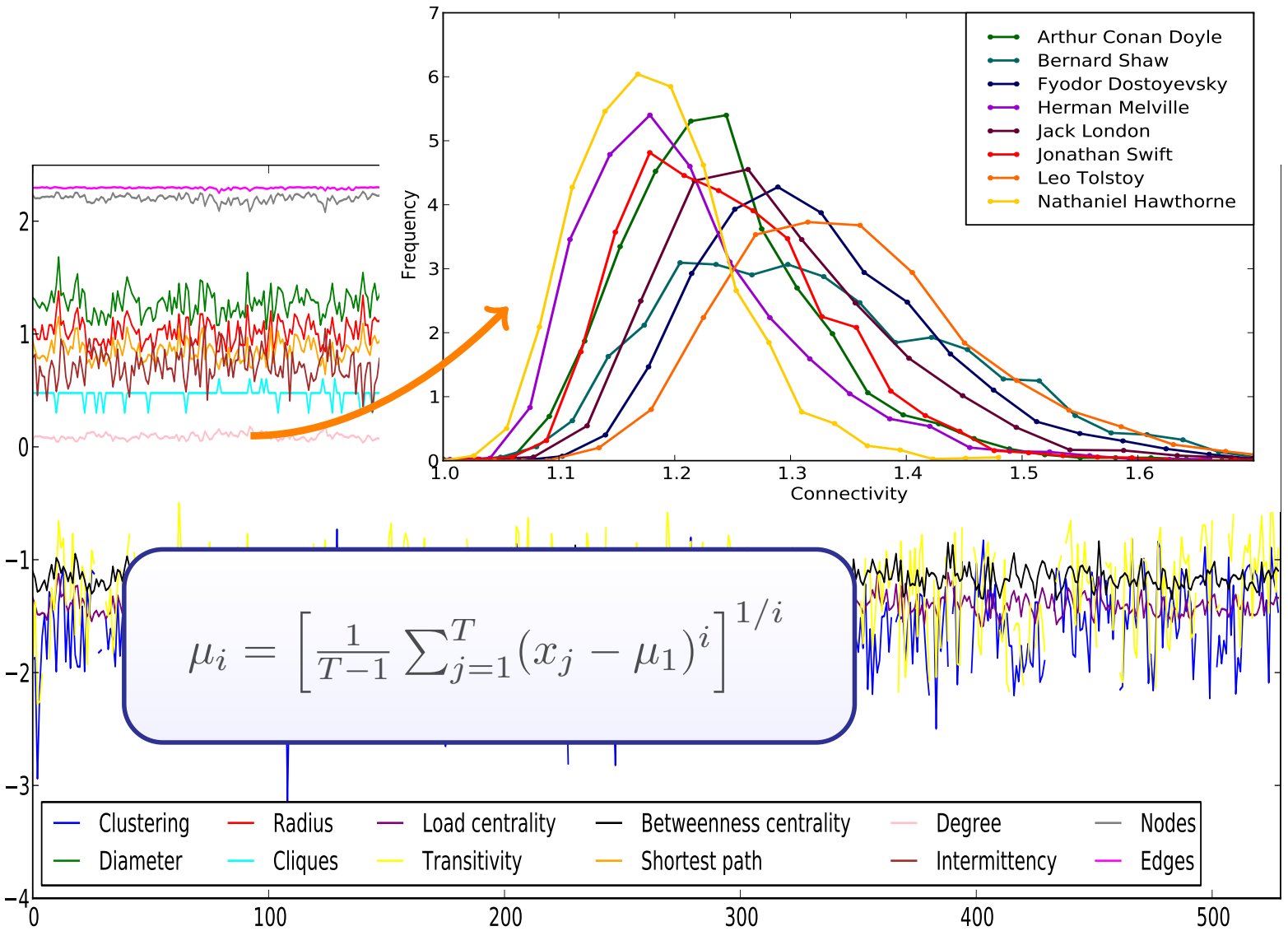
Time series for Moby Dick by H. Melville

Time series



Time series for Moby Dick by H. Melville

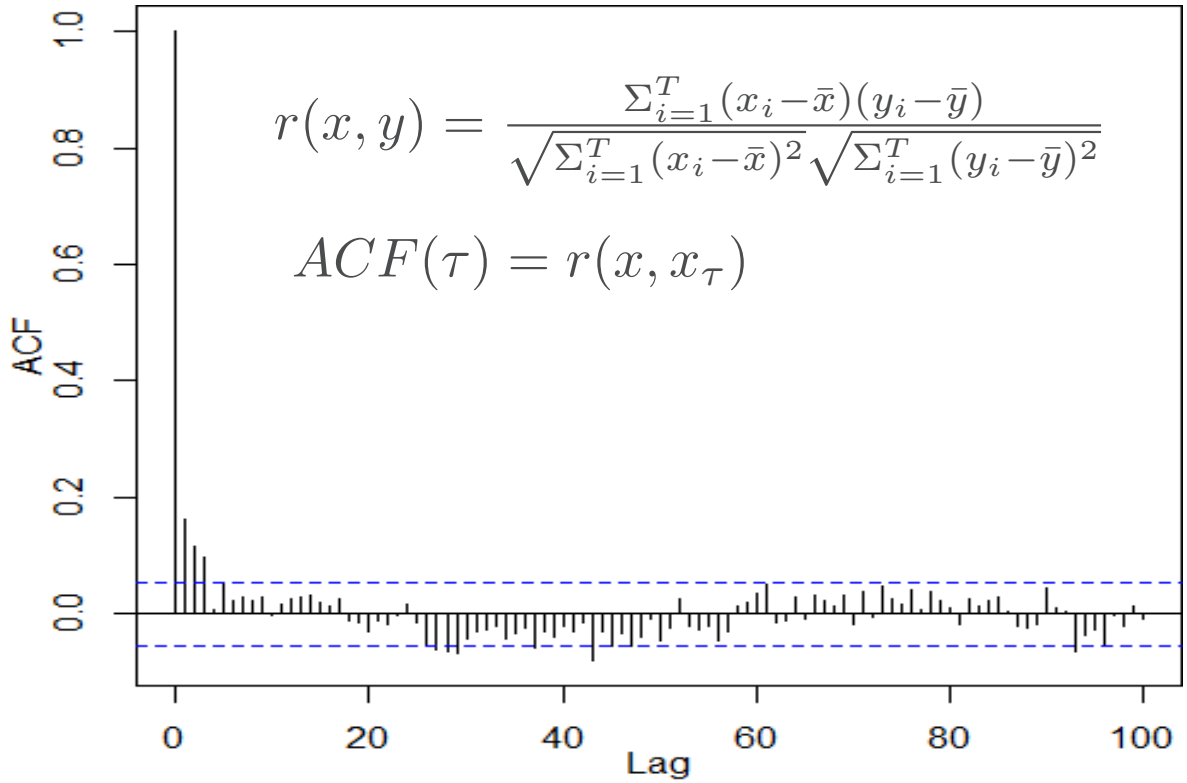
Time series



Time series for Moby Dick by H. Melville

Time series

Autocorrelation



$$r(x,y) = \frac{\Sigma_{i=1}^{T}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma_{i=1}^{T}(x_i-\bar{x})^2}\sqrt{\Sigma_{i=1}^{T}(y_i-\bar{y})^2}}$$

$$ACF(\tau) = r(x, x_\tau)$$

$\pm 5\%$

Wiener-Khinchin theorem

$$C(\tau) \equiv \int_{-\infty}^{\infty} x^*(t)x(\tau+t)\mathrm{d}t = \int_{-\infty}^{\infty} |x_\nu|^2 e^{-2\pi\imath\nu\tau}\mathrm{d}\nu = \mathcal{F}[|x_\nu|^2](\tau)$$

Time series

Stationarity tests and ARIMA fittings

Auto-regressive model AR(p)

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + \varepsilon_t, \quad t > p$$

Characteristic equation: $1 - a_1 z - a_2 z^2 + \cdots + a_p z^p = 0$
Unit root tests: $z = 1$?

Auto-Regressive Integrated Moving Average model ARIMA(p,d,q)

$$\left(1 - \sum_{i=0}^{p} \phi_i L^i\right) (1 - L)^d x_t = \left(1 + \sum_{i=0}^{q} \theta_i L^i\right) \varepsilon_t$$

Lag operator: $L x_t = x_{t-1}$

Time series

Stationarity tests

	Phillips-Perron		KPSS	Dickey-Fuller	McKinnon	
Clustering	0.010		0.071	0.017	0.008	0.167
Betweenness centrality	0.023		0.074	0.350	0.360	0.510
Cliques	0.010		0.086	0.377	0.393	0.521
Diameter	0.010		0.076	0.116	0.111	0.365
Intermittency	0.010		0.071	0.080	0.074	0.335
Load centrality	0.081		0.080	0.457	0.478	0.583
Degree	0.019		0.066	0.470	0.513	0.579
Radius	0.011		0.073	0.118	0.114	0.368
Shortest path	0.013		0.071	0.214	0.208	0.430
Edges	0.253		0.078	0.362	0.369	0.512
Nodes	0.022		0.067	0.368	0.378	0.514
Transitivity	0.010		0.083	0.014	0.005	0.126

$p_{value} > 0.05 \quad p_{value} < 0.05$

Time series

ARIMA fittings

Network metric	Value of d		
	0	1	2
Clustering	55	25	0
Betweenness centrality	57	23	0
Cliques	69	11	0
Diameter	60	20	0
Intermittency	56	24	0
Load centrality	63	17	0
Degree	51	29	0
Radius	58	22	0
Shortest path	55	25	0
Edges	61	19	0
Nodes	49	31	0
Transitivity	64	16	0
Total	698	262	0

ARIMA(p,0,q)

Stationary

73%

ARIMA(p,1,q)

First order integrated

27%

Time series

ARIMA fittings

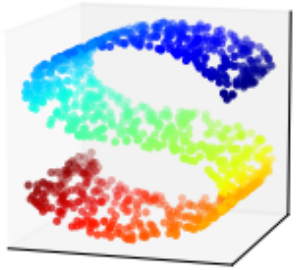
Table: Series fitted with an ARIMA(p,d,q) model having the biggest values of the sum $p + d + q$.

			ARIMA(p,d,q)		
Book	Measure	Sum	p	d	q
The Poems of Jonathan Swift, D.D., Volume 2	Load centrality	9	5	0	4
The Journal to Stella	Clustering	8	2	1	5
The Iron Heel	Clustering	8	3	1	4
Typee: A Romance of the South Seas	Edges	8	4	1	3

Data analysis

Dimensionality reduction

- Feature selection
- Feature extraction



48 Attributes

80 Books

8 Authors

Supervised learning

- Zero Rule: $1/8 = 12.5\%$
- One Rule
- Naive Bayes
- K-Nearest Neighbors
- J48 (tree)
- Radial Basis Function Networks

Precision

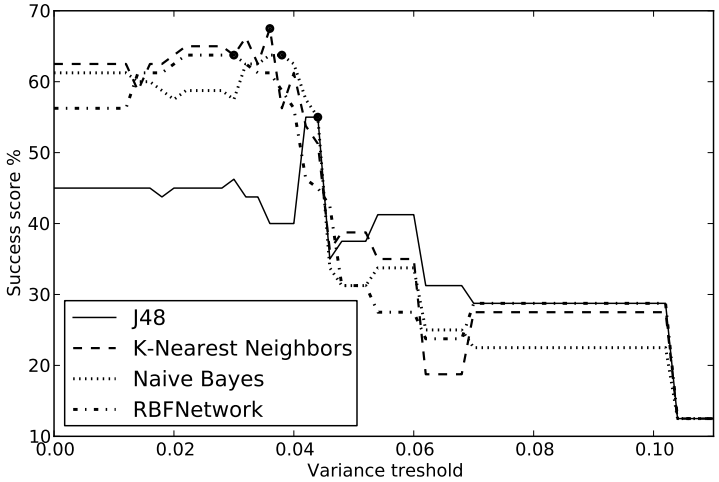
$$P_A = \frac{TP_A}{TP_A + FP_A}$$

Recall

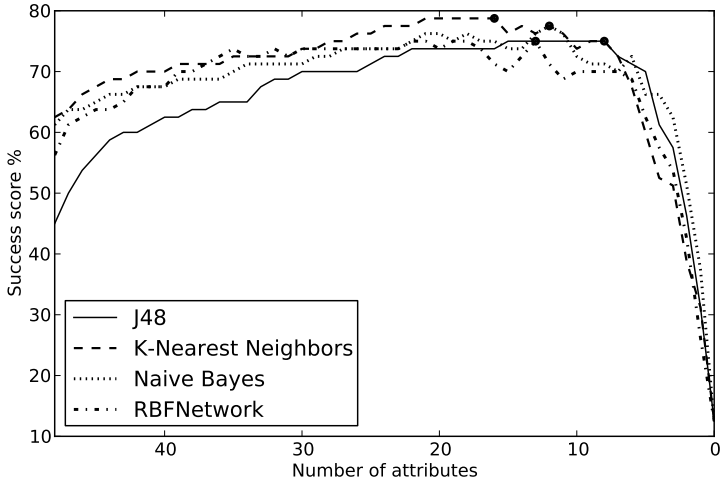
$$R_A = \frac{TP_A}{TP_A + FN_A}$$

TP: True Positives FP: False Positives
FN: False Negatives

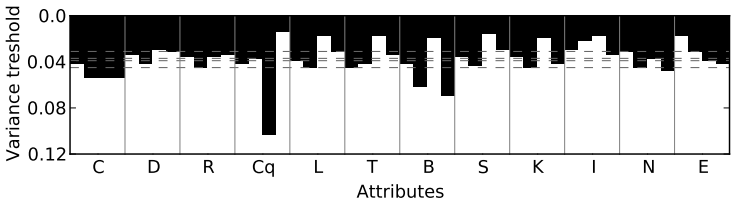
Feature selection



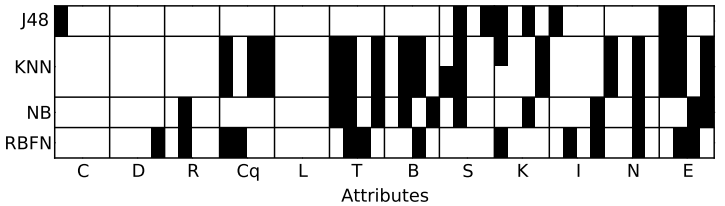
Success scores using variance threshold.



Success scores using score-based criteria.



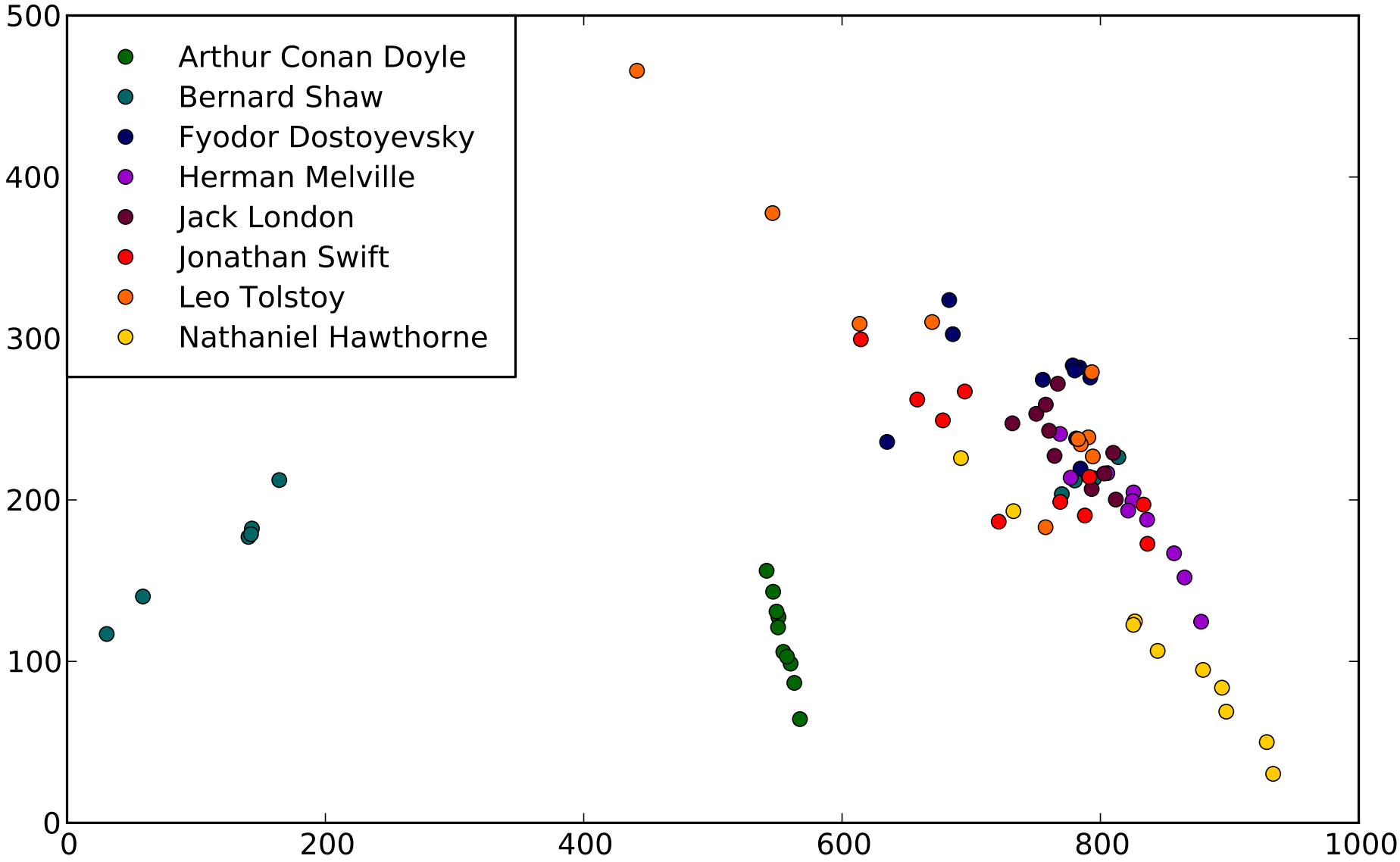
Features using variance threshold.



Features using score-based criteria.

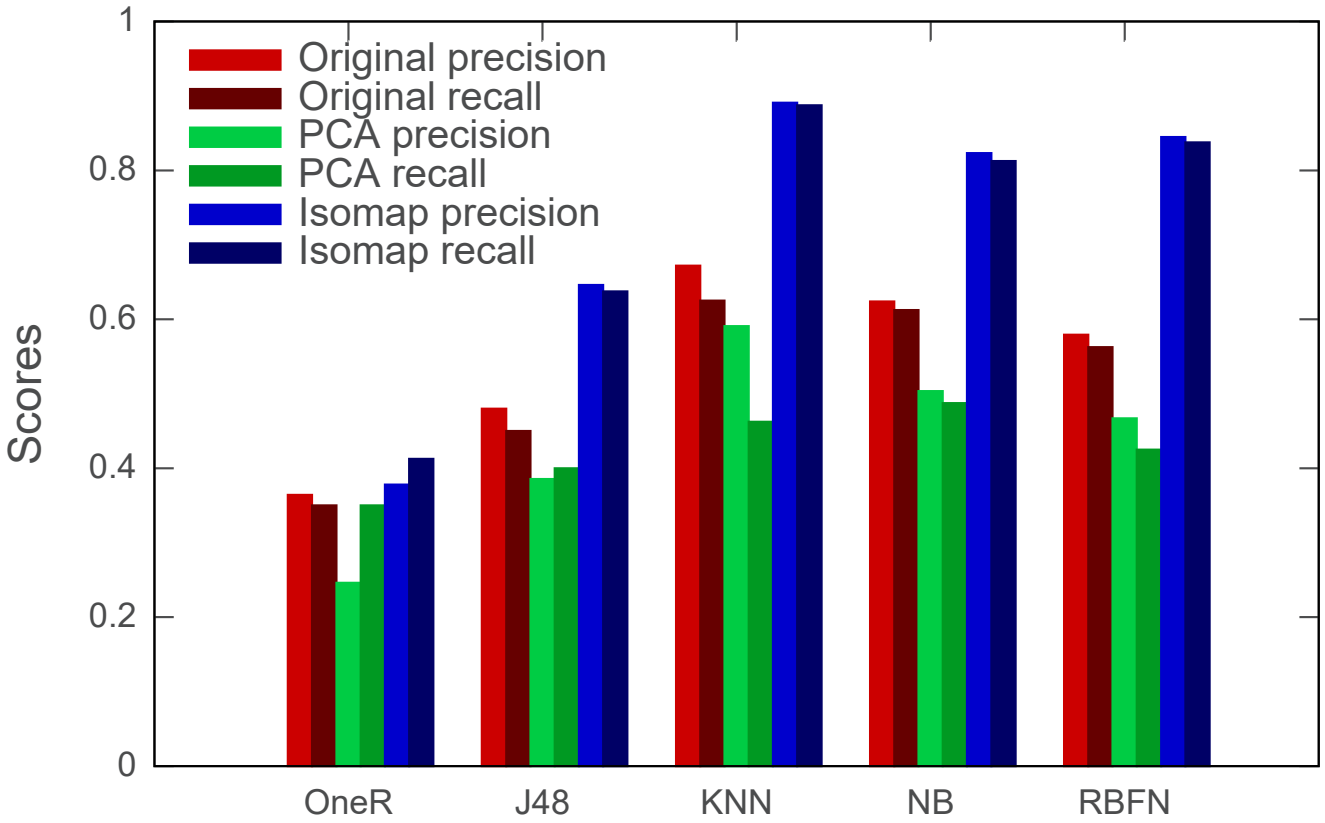
Success scores and combinations of features using feature selection. In the upper figures maximum values are marked with circles. In the lower figures if an attribute is present in the combination the corresponding cell is painted black.

Feature extraction



Feature extraction using ISOMAP.

Feature extraction



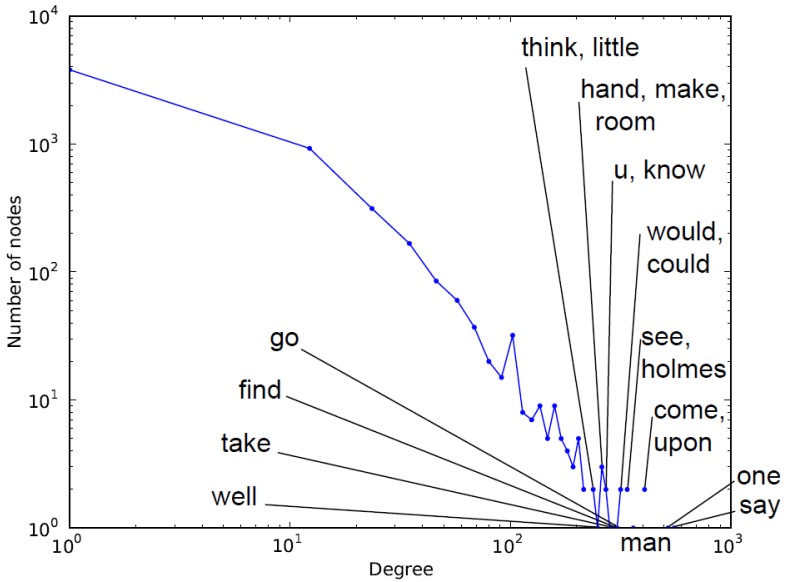
Scores using feature extraction.

Summary of classification success scores

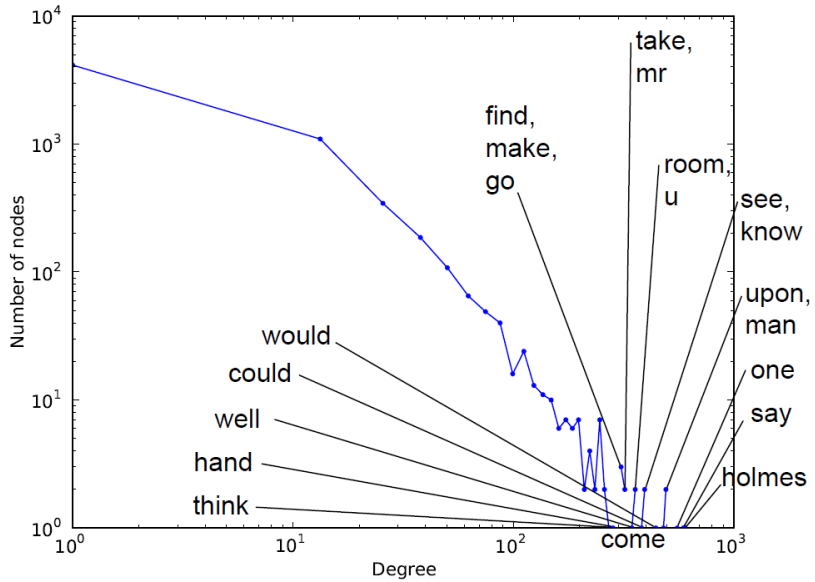
Attributes	**J48** (%)	**KNN** (%)	**NB** (%)	**RBFN** (%)
Original set	45.00	62.50	61.25	56.25
Variance threshold best	55.00	67.50	63.75	63.75
Score-based best	75.00	78.75	77.50	75.00
$\{\mu_1\}$	45.00	43.75	46.25	40.00
$\{\mu_2, \mu_3, \mu_4\}$	38.75	63.75	60.00	57.50
PCA	40.00	46.25	48.75	42.50
ISOMAP	63.75	88.75	81.25	83.75

The role of words

The Memoirs of Sherlock Holmes

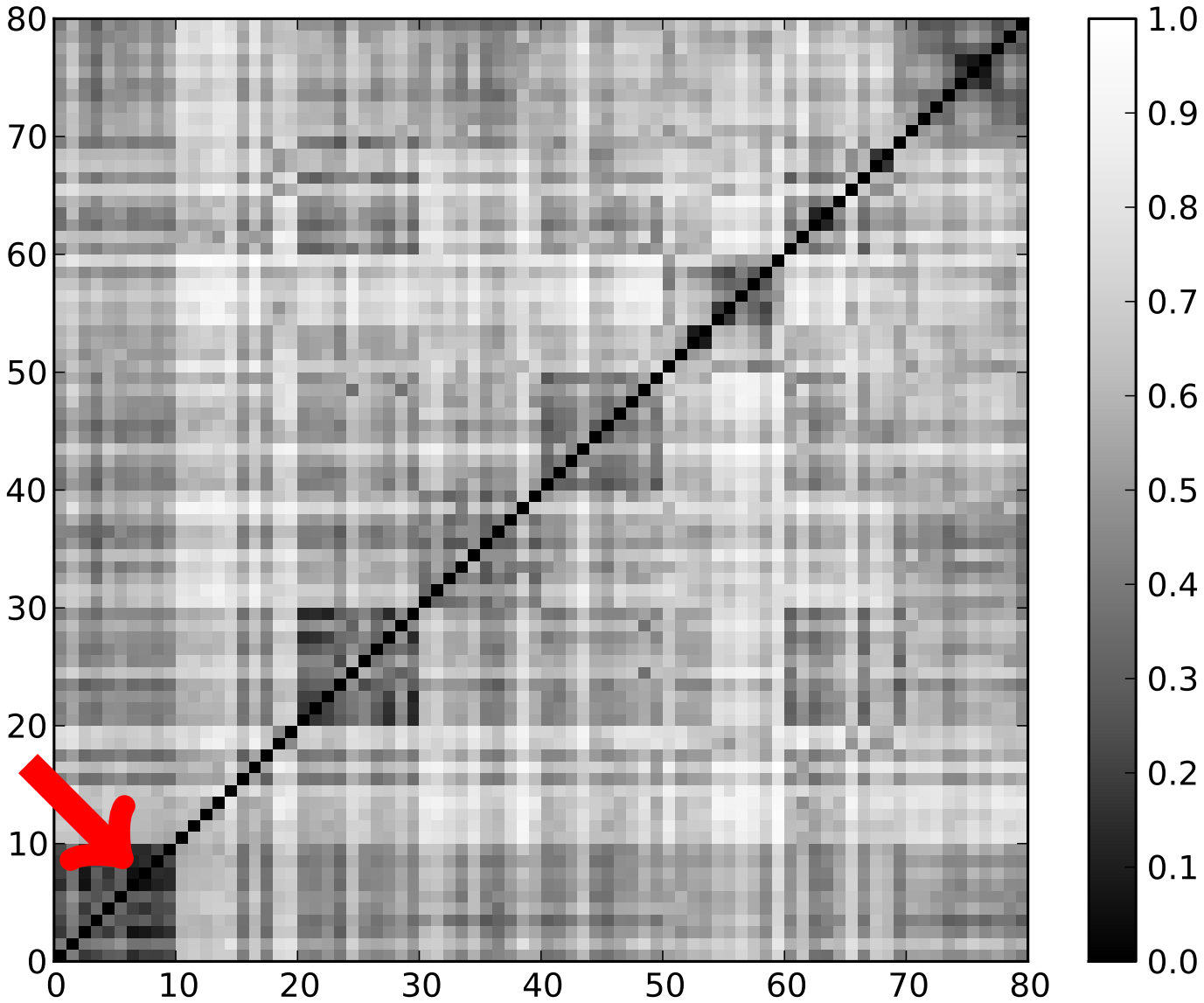


The Return of Sherlock Holmes

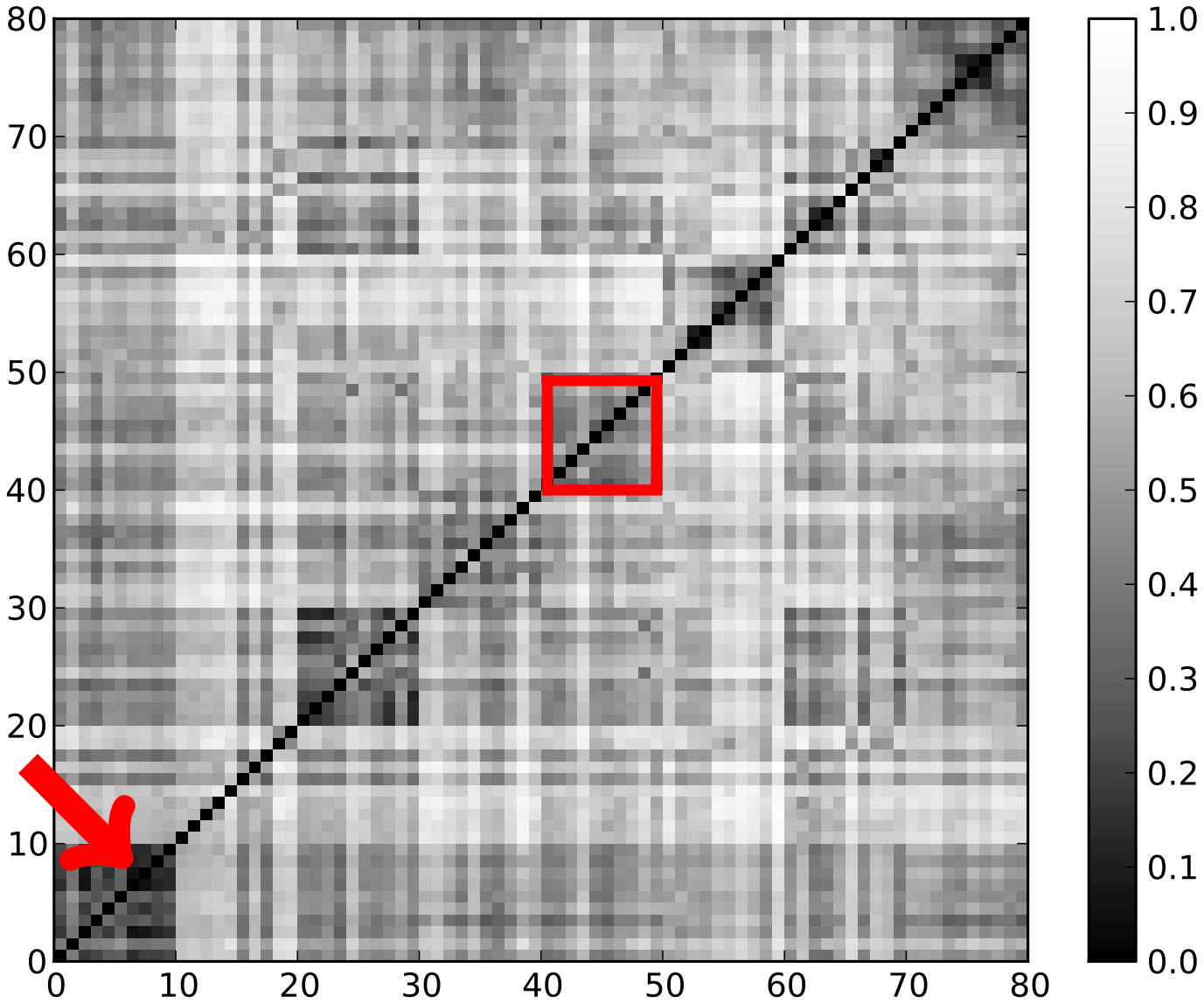


Only one different word out of the 20 highest ranked!

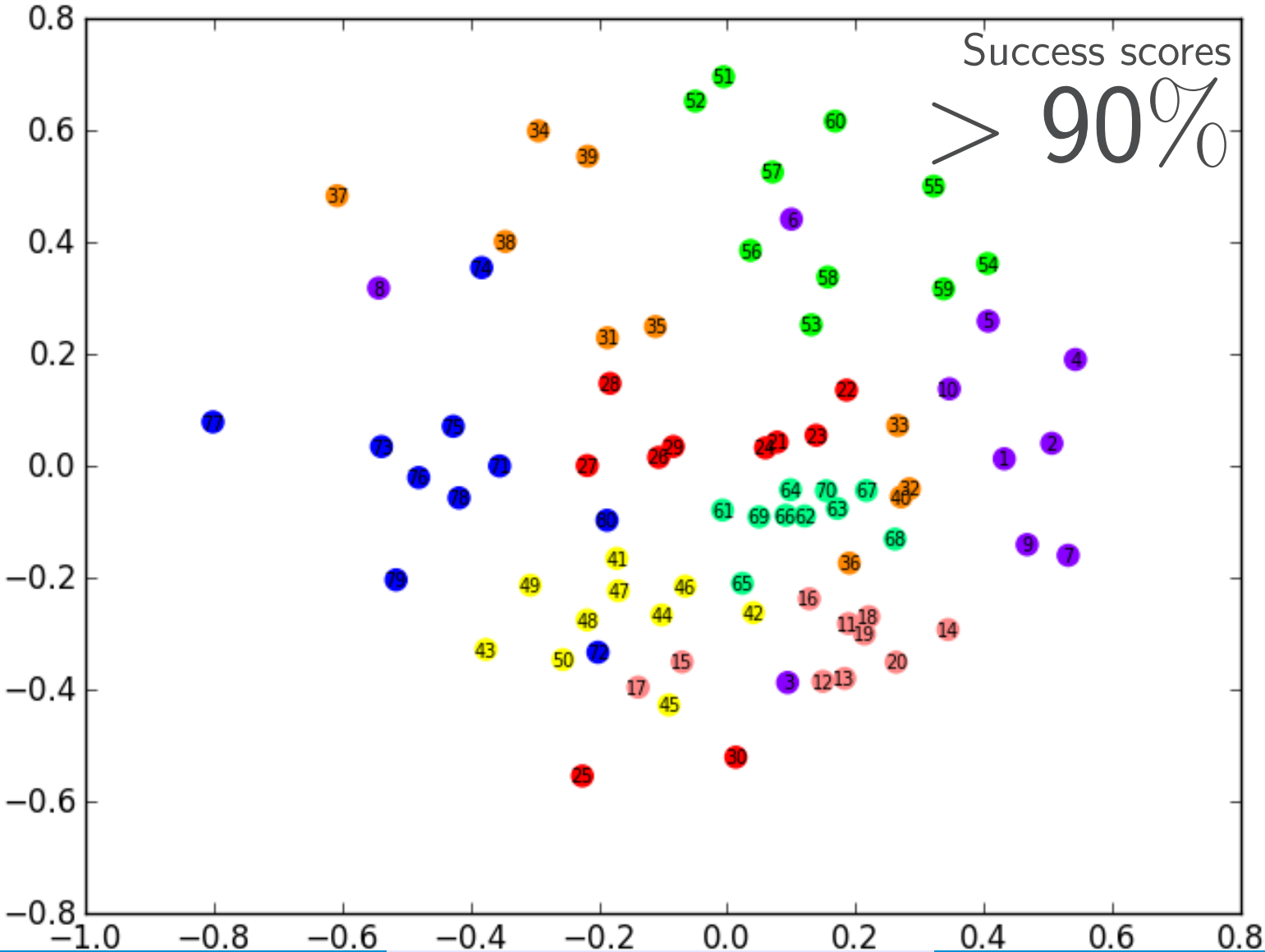
Dissimilarity matrix



Dissimilarity matrix



Projection


Success scores
> 90%
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

Summary

- Time series are stationary.
- Global sample statistics can be obtained.
- Dynamic measures are author-dependent.
- Weight on edges is relevant.
- Dimensionality reduction enhances classification.
- Books are located on a curved manifold in attribute space.
- A word's role in a network is author-dependent.
- Network metrics must be jointly used for classification.
- Many hidden features of networks.

Muchas gracias!