Networks based on words

Bowen Dai

WANs definition

- Word-adjacency networks belong to the large class of word co-occurrence networks

- Given a set of words W and a list of k corpora $C=\{c_1,c_2...c_k\}$, the undirected co-occurrence network is defined as $G=\{W,E(W,C)\}$ where $\{w_1, w_2\} \in E(W,C)$ if w_i and w_j co-occur in at least on corpus.

The small word of human language

- The so called small-world effect. In particular, the average distance between two words, d (i.e. the average minimum number of links to be crossed from an arbitrary word to another), is shown to be d º 2^3, even though the human brain can store many thousands

The small word of human language

- A scale-free distribution of degrees
- A scale-free network is a network whose degree distribution follows a power law, at least asymptotically.

The small word of human language

- lexicon kernel
- co-occurrence of words in sentences relies on the network structure of the lexicon

The small word of human language

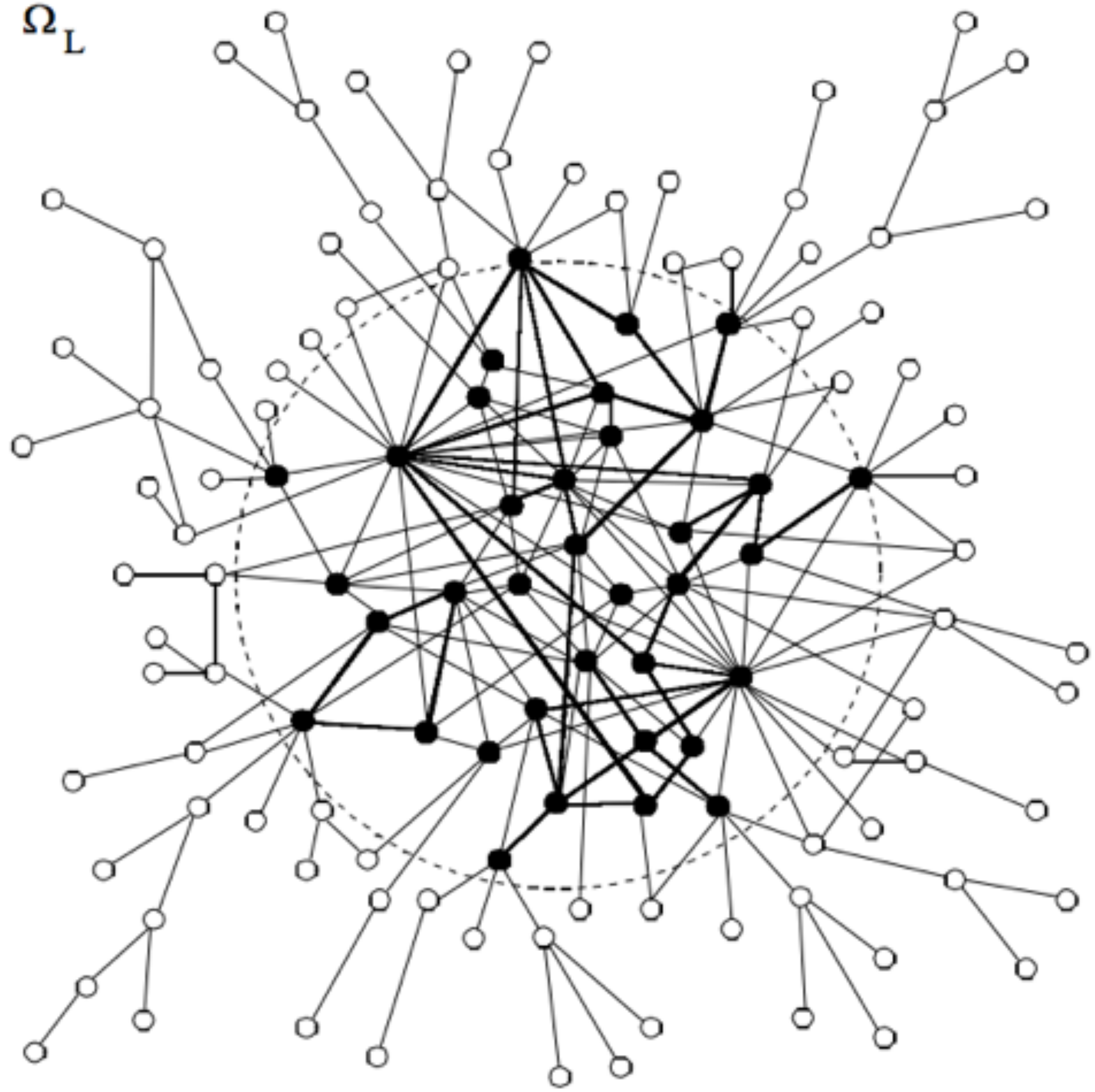
- For random graphs, $C_{\mathrm{v}}^{\mathrm{rand}} \approx \bar{k}/N$.
- For SW graphs, d is close to that expected for random graphs, d^{rand}, with the same k and $C_{\mathrm{v}} \gg C_{\mathrm{v}}^{\mathrm{rand}}$.
- These two conditions are taken as the standard de¢nition of SW

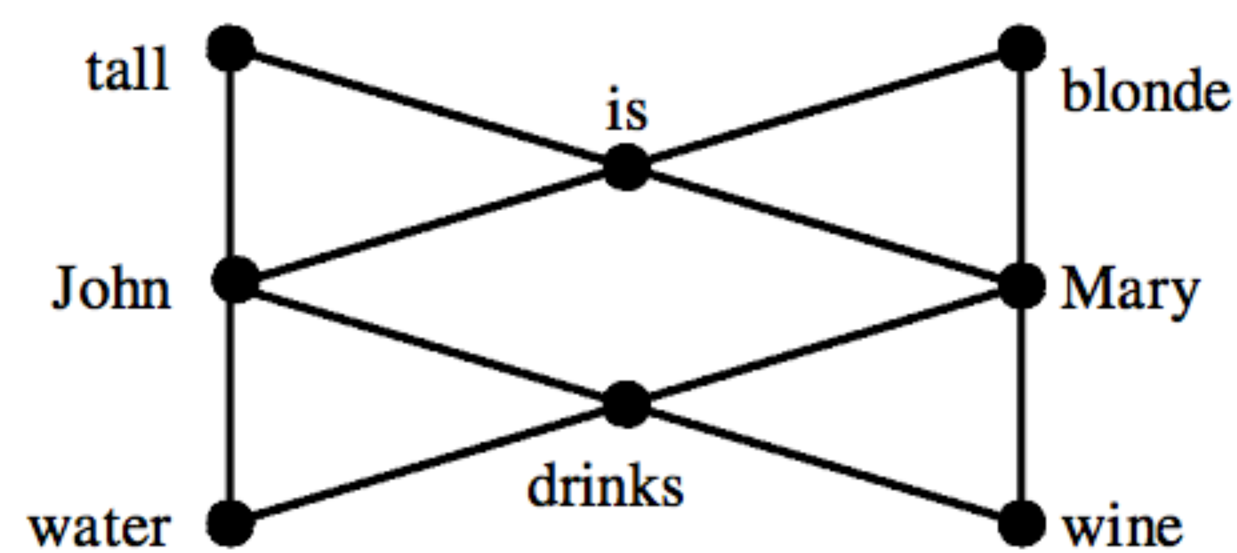
Table 1. Word network patterns.

(It can be seen that $C \gg C_{\mathrm{random}}$ and $d \approx d_{\mathrm{random}}$, consistently in a SW network. All values are exact except for those marked with an asterisk, which have been estimated on a random subset of the vertices (after having processed 2% of the vertices, fluctuations in d^* as a function of the subset size clearly affected only the third decimal digit).)

graph	C	C_{random}	d	d_{random}
Ω_L (UWN)	0.687	1.55×10^{-4}	2.63*	3.03
Ω_L (RWN)	0.437	1.55×10^{-4}	2.67*	3.06

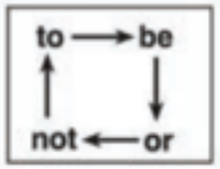
The small word of human language

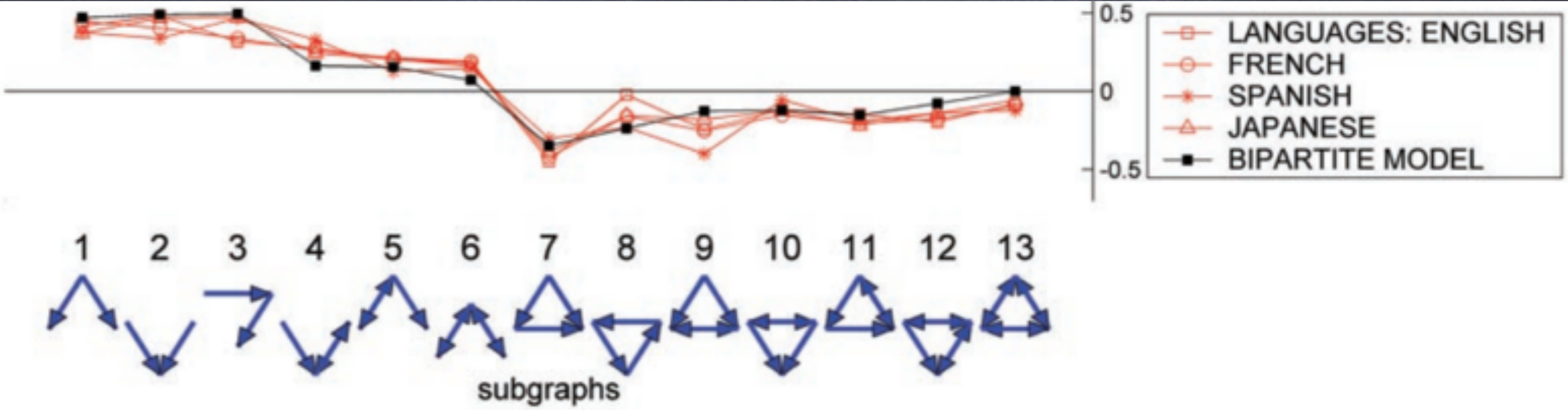

Ω_L


tall
is
blonde
John
Mary
drinks
water
wine

Triad significance profile

The TSP shows the normalized significance level (Z score) for each of the 13 triads





Application of WANs

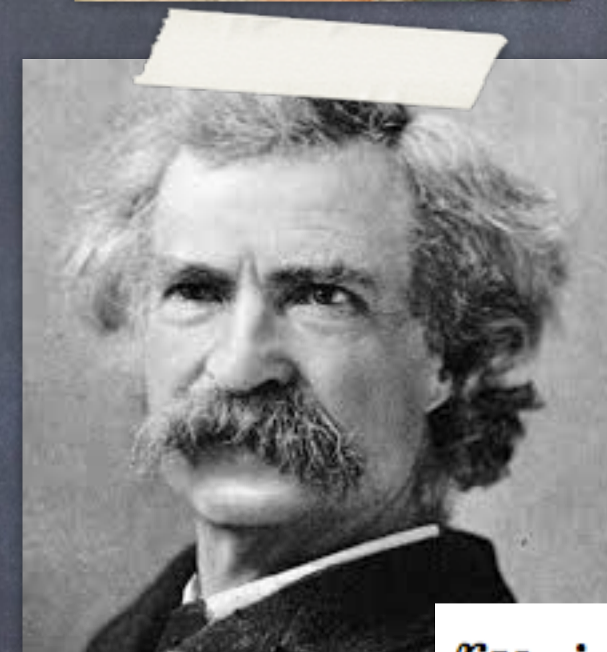
Authorship Attribution

- Encode structures as word adjacency networks (WANs) which are asymmetric networks that store information of co-appearance of two function words in the same sentence
- With proper normalization, edges of these networks describe the likelihood that a particular function word is encountered in the text given that we encountered another one. In turn, this implies that WANs can be reinterpreted as Markov chains describing transition probabilities between function words.

Northanger Abby
Sense and Sensibility
Pride and Prejudice

Emma

The Adventures of Tom Sawyer
A Connecticut Yankee in King Arthur's Court
The Innocents Abroad



Eve's Diary

$r_T : T \to A$

$r_U : U \to A$

Bartleby, the Scrivener
Typee
Omoo

Redburn

Authorship Attribution

- For a given sentence, we define a directed proximity between two words parametric on a discount factor $\alpha \in (0, 1)$ and a window length D. If we denote as $i(\omega)$ the position of word ω within its sentence the directed proximity $d(\omega 1, \omega 2)$ from word $\omega 1$ to word $\omega 2$ when $0 < i(\omega 2) - i(\omega 1) \leq D$ is defined as

$$d(\omega_1, \omega_2) := \alpha^{i(\omega_2) - i(\omega_1) - 1}.$$

Authorship Attribution

- both w1 and w2 are function words

Common Function Words									
the	and	a	of	to	in	that	with	as	it
for	but	at	on	this	all	by	which	they	so
from	no	or	one	what	if	an	would	when	will

Authorship Attribution

- parameter $\alpha = 0.8$, the window $D = 4$
- <u>a</u> swarm <u>in</u> May is worth <u>a</u> load <u>of</u> hay; <u>a</u> swarm <u>in</u> June is worth <u>a</u> silver spoon; <u>but</u> <u>a</u> swarm <u>in</u> July is not worth <u>a</u> fly
- a swarm in May is worth a load of hay
- a swarm in June is worth a silver spoon
- but a swarm in July is not worth a fly

Authorship Attribution

- Function WANs
- function words as nodes
- The weight of a given edge represents the likelihood of finding the words connected by this edge close to each other in the text
- from a given text t we construct the network $W_t = (F, Q_t)$ where $F = \{f_1, f_2, \ldots, f_f\}$ is the set of nodes composed by a collection of function words common to all WANs being compared and $Q_t : F \times F \rightarrow R^+$ is a similarity measure between pairs of nodes.

Authorship Attribution

$$Q_t(f_i, f_j) = \sum_{h,e} \mathbb{I}\{s_t^h(e) = f_i\} \sum_{d=1}^{D} \alpha^{d-1}\, \mathbb{I}\{s_t^h(e+d) = f_j\},$$

- s(e) is the word in the e-th position within sentence h of text t

$$Q_t = \begin{array}{c} \\ \text{a} \\ \text{in} \\ \text{of} \\ \text{but} \end{array} \begin{array}{c} \begin{array}{cccc} \text{a} & \text{in} & \text{of} & \text{but} \end{array} \\ \begin{pmatrix} 0 & 3 \times 0.8^1 & 0.8^1 & 0 \\ 2 \times 0.8^3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0.8^2 & 0 & 0 \end{pmatrix} \end{array}.$$

Authorship Attribution

$$Q_c = \sum_{t \in T^{(c)}} Q_t.$$

$$\hat{Q}_c(f_i, f_j) = \frac{Q_c(f_i, f_j)}{\sum_j Q_c(f_i, f_j)},$$

- sum all matrix for the same author and then create the markov chain

$$\hat{Q}_t = \begin{array}{c c} & \begin{array}{cccc} \text{a} & \text{in} & \text{of} & \text{but} \end{array} \\ \begin{array}{c} \text{a} \\ \text{in} \\ \text{of} \\ \text{but} \end{array} & \left(\begin{array}{cccc} 0 & 0.75 & 0.25 & 0 \\ 1 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.61 & 0.39 & 0 & 0 \end{array} \right) \end{array}.$$

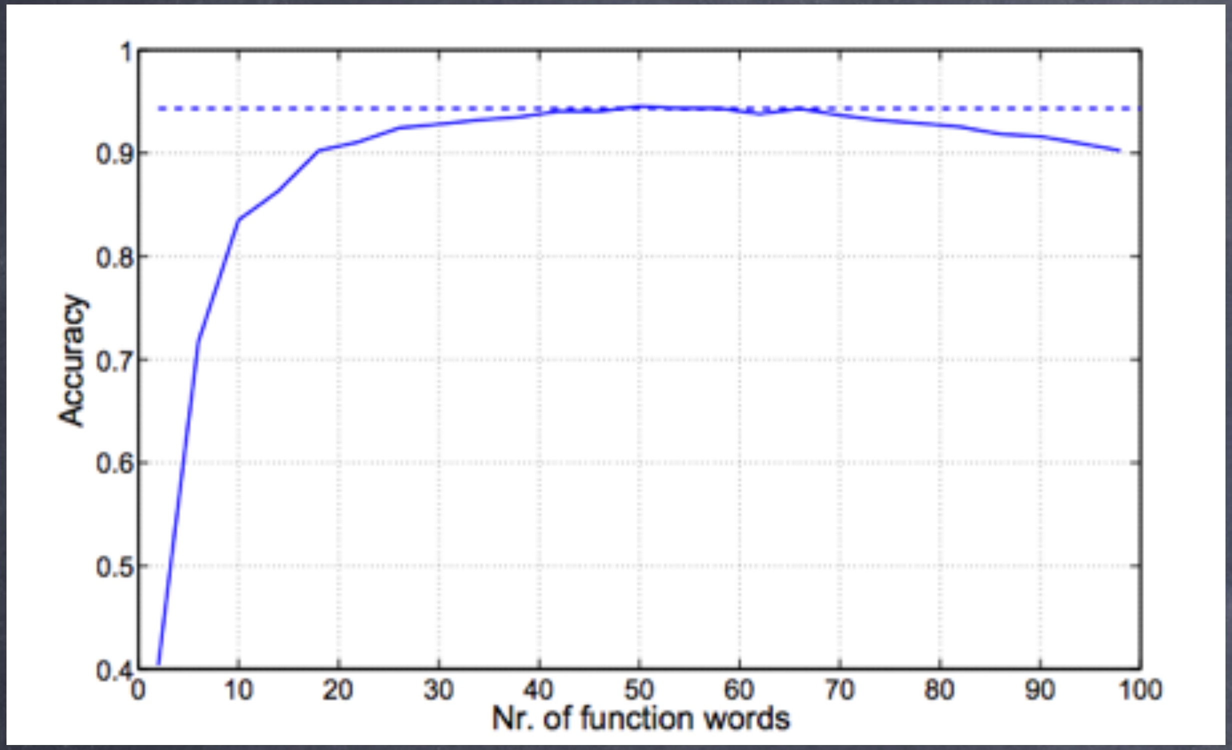
Authorship Attribution

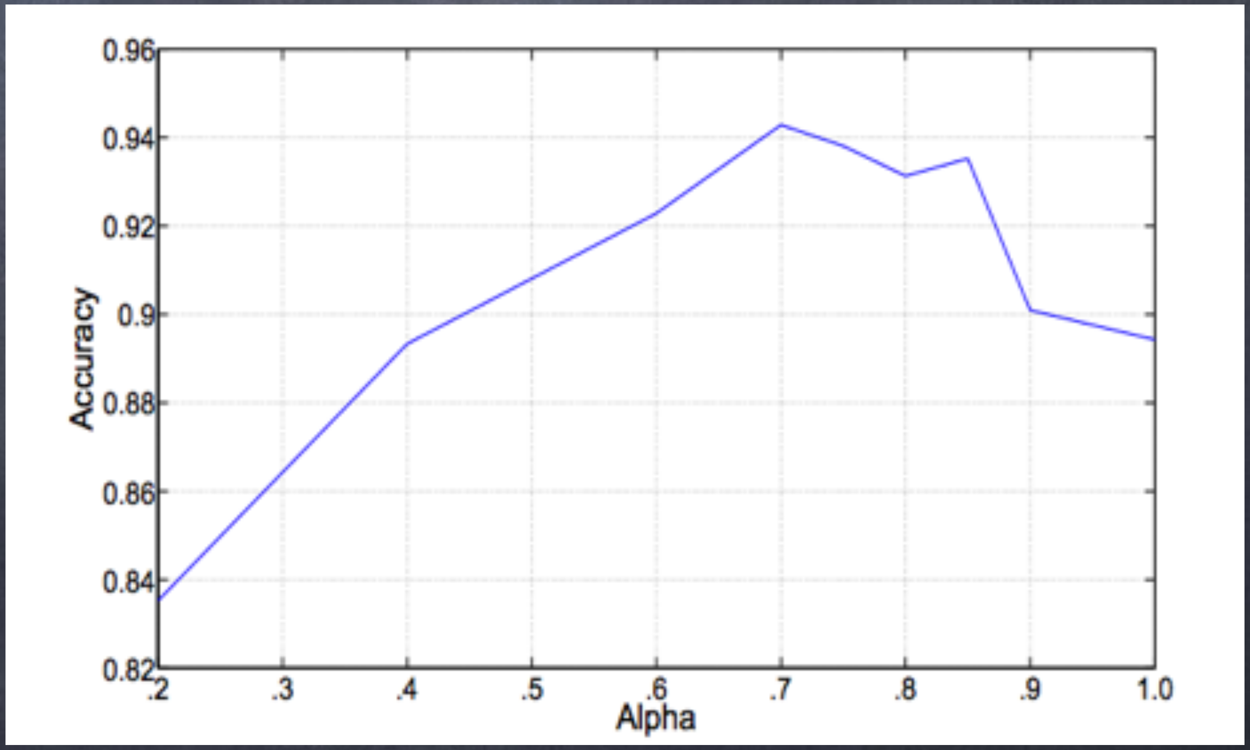
- The normalized networks P can be interpreted as discrete time Markov chains (MC)
- Since every MC has the same state space F, we use the relative entropy H(P1, P2) as a dissimilarity measure between the chains P1 and P2. The relative entropy is given by

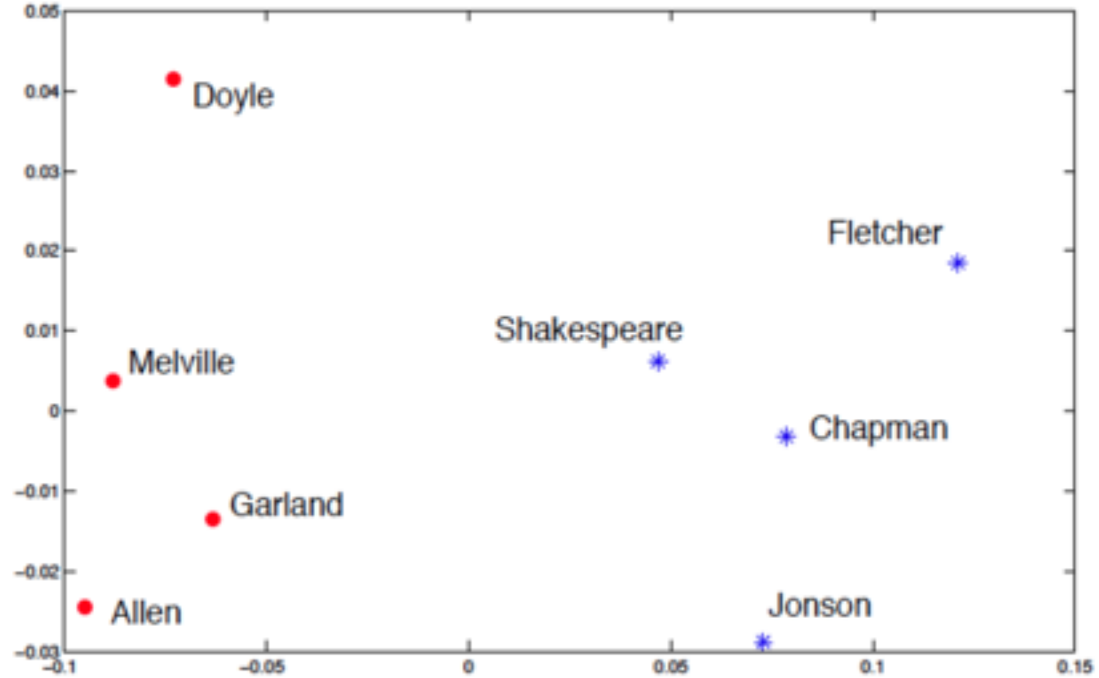
$$H(P_1, P_2) = \sum_{i,j} \pi(f_i) P_1(f_i, f_j) \log \frac{P_1(f_i, f_j)}{P_2(f_i, f_j)},$$

$$H(P_1, P_2) = \sum_{i,j | P_2(f_i, f_j) \neq 0} \pi(f_i) P_1(f_i, f_j) \log \frac{P_1(f_i, f_j)}{P_2(f_i, f_j)}.$$

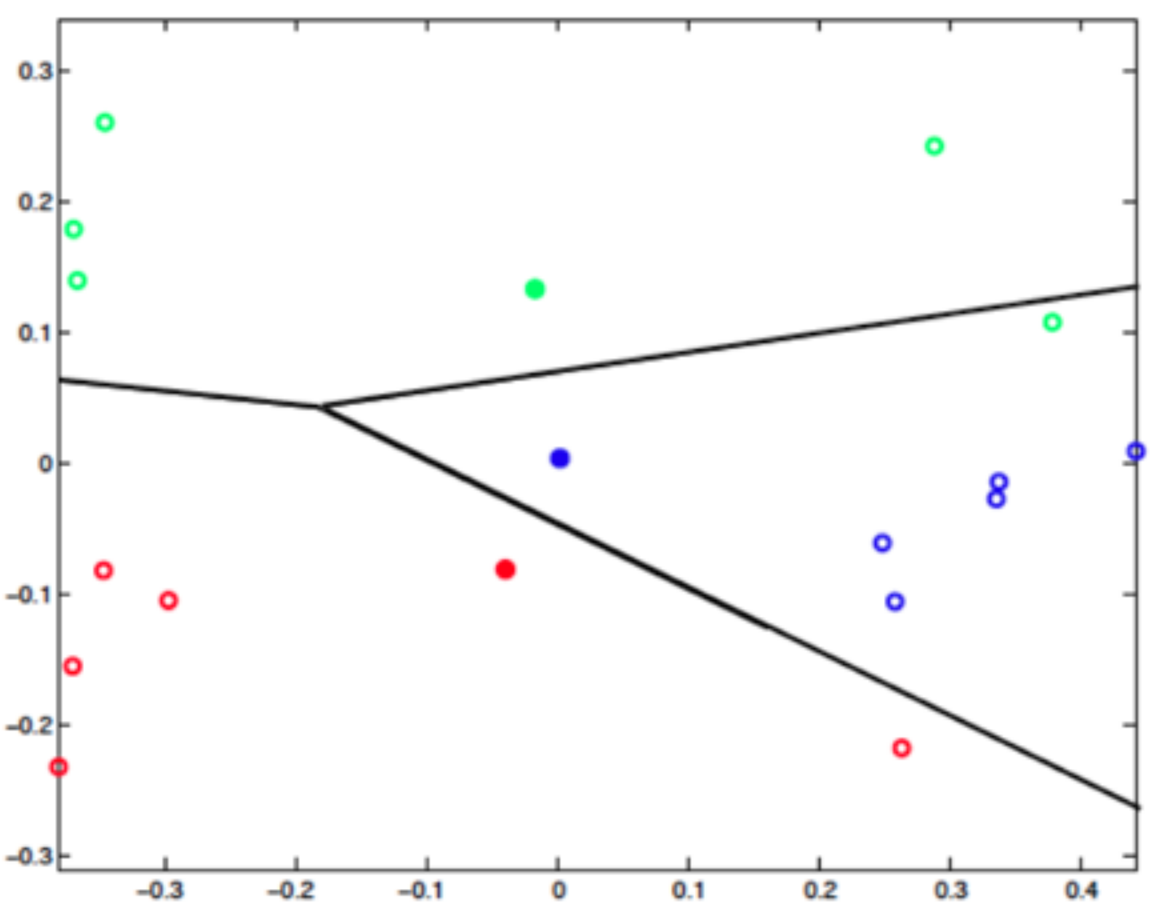
Authorship Attribution







(a) MDS plot for authors of different time periods.



(b) MDS representation for three authors.

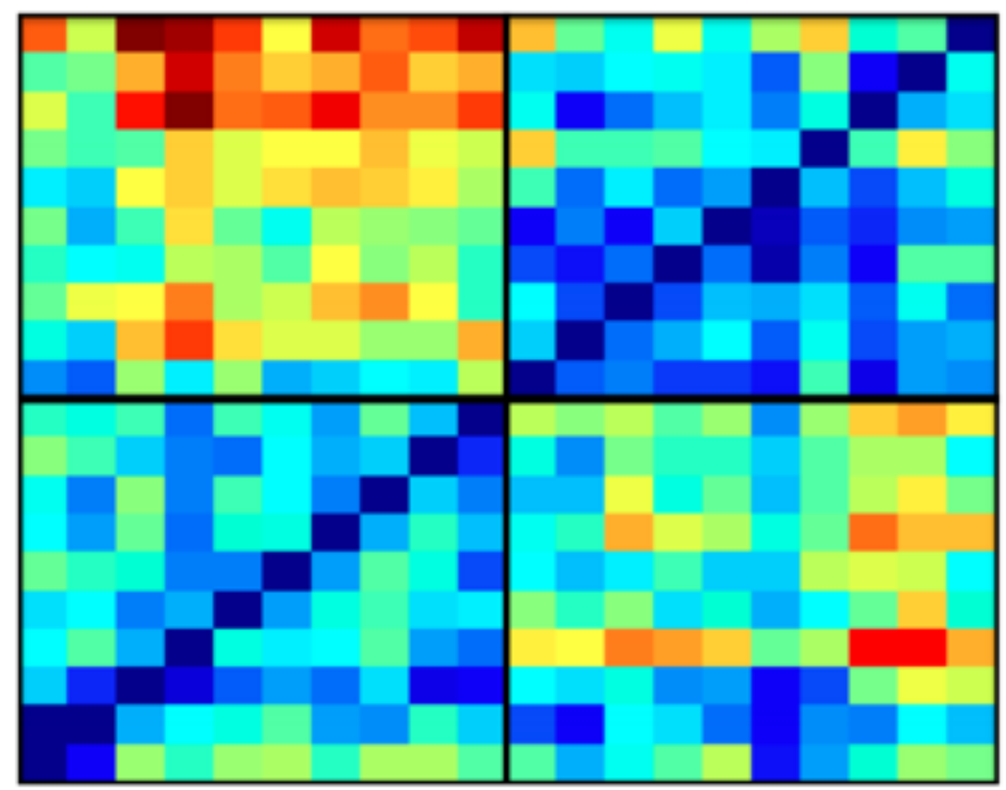


Fig. 5: Heat map of relative entropies between 20 Shakespeare extracts. The first 10 texts correspond to history plays while the last 10 correspond to comedy plays. Relative entropies within texts of the same genre are smaller than across genres.

Future Topic

- What's next after we find a network satisfied SW
- Markov chain
- dai.171@osu.edu

bibliography

- Segarra, S., Eisen, M., & Ribeiro, A. (2015). Authorship attribution through function word adjacency networks. Signal Processing, IEEE Transactions on, 63(20), 5464-5478.
- Ferrer, I. C. R., & Solé, R. V. (2001, November). The small world of human language. In Proceedings. Biological sciences/The Royal Society (Vol. 268, No. 1482, pp. 2261-2265).
- Zweig, K. A. (2016). Are Word-Adjacency Networks Networks?. In Towards a Theoretical Framework for Analyzing Complex Linguistic Networks (pp. 153-163). Springer Berlin Heidelberg.

bibliography

- Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., ... & Alon, U. (2004). Superfamilies of evolved and designed networks. Science, 303(5663), 1538-1542.
- Choudhury, M., Chatterjee, D., & Mukherjee, A. (2010, August). Global topology of word co-occurrence networks: Beyond the two-regime power-law. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 162-170). Association for Computational Linguistics.
- Choudhury, M., & Mukherjee, A. (2009). The structure and dynamics of linguistic networks. In Dynamics on and of Complex Networks (pp. 145-166). Birkhäuser Boston.