## Authorship recognition

<table>
<thead>
<tr>
<th><strong>Lexical features (frequency)</strong></th>
<th><strong>Text format</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Words, $n$-grams, functional words, types of words, discourse-connecting expressions, slang, contractions, dialects, orthography mistakes, proper names, semantic features (polysemy).</td>
<td>Lengths of lines, words and phrases, formatting (white spaces), capitalization, non-alphanumeric characters, beginnings and ends of texts.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Character-level features</strong></th>
<th><strong>Other</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Character $n$-grams, frequent suffixes, punctuation.</td>
<td>Syntactic features: $n$-grams syntactic function, kinds of phrases, perplexity, morphological complexity.</td>
</tr>
</tbody>
</table>
Complex networks

Natural

Artificial

Social

Other

Ocean

World
Complexity of language

Zipf’s law (1939)

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

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


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
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

C. Akimushkin (IFSC-USP)
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: \( T = \frac{3 \text{triangles}}{\text{triads}} \)
7. Betweenness centrality: \( B_i = \sum_{s \neq i \neq t} \frac{g_{st}^i}{g_{st}} \)
8. Shortest path: \( l_{ij} = [A^n]_{ij} \)
9. Connectivity: \( k_i = [A^2]_{ii} \)
10. Intermittency: \( I_i = \text{var}(\Delta)/\bar{\Delta} \)
11. Number of nodes: \( N \)
12. Number of edges: \( E \)

A: Adjacency matrix; \( g_{st} = \sum l_{st}; \) \( \Delta_i \): 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 for Moby Dick by H. Melville

- Clustering
- Radius
- Load centrality
- Betweenness centrality
- Degree
- Diameter
- Clique
- Transitivity
- Shortest path
- Intermittency
- Nodes
- Edges
Time series for Moby Dick by H. Melville

- Arthur Conan Doyle
- Bernard Shaw
- Fyodor Dostoyevsky
- Herman Melville
- Jack London
- Jonathan Swift
- Leo Tolstoy
- Nathaniel Hawthorne

Time series

Connectivity

Clustering
Diameter
Radius
Cliques
Load centrality
Transitivity
Betweenness centrality
Shortest path
Degree
Intermittency
Nodes
Edges

\[ \mu_i = \frac{1}{T-1} \sum_{j=1}^{T} (x_j - \mu_1) \]
Time series

\[ \mu_i = \left[ \frac{1}{T-1} \sum_{j=1}^{T} (x_j - \mu_1)^i \right]^{1/i} \]

Time series for Moby Dick by H. Melville
Time series

Autocorrelation

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

\[ ACF(\tau) = r(x, x_\tau) \]

Wiener-Khinchin theorem

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

<table>
<thead>
<tr>
<th></th>
<th>Phillips-Perron</th>
<th>KPSS</th>
<th>Dickey-Fuller</th>
<th>McKinnon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td>0.010</td>
<td>0.071</td>
<td>0.017</td>
<td>0.008</td>
</tr>
<tr>
<td>Betweenness centrality</td>
<td>0.023</td>
<td>0.074</td>
<td>0.350</td>
<td>0.360</td>
</tr>
<tr>
<td>Clique</td>
<td>0.010</td>
<td>0.086</td>
<td>0.377</td>
<td>0.393</td>
</tr>
<tr>
<td>Diameter</td>
<td>0.010</td>
<td>0.076</td>
<td>0.116</td>
<td>0.111</td>
</tr>
<tr>
<td>Intermittency</td>
<td>0.010</td>
<td>0.071</td>
<td>0.080</td>
<td>0.074</td>
</tr>
<tr>
<td>Load centrality</td>
<td>0.081</td>
<td>0.080</td>
<td>0.457</td>
<td>0.478</td>
</tr>
<tr>
<td>Degree</td>
<td>0.019</td>
<td>0.066</td>
<td>0.470</td>
<td>0.513</td>
</tr>
<tr>
<td>Radius</td>
<td>0.011</td>
<td>0.073</td>
<td>0.118</td>
<td>0.114</td>
</tr>
<tr>
<td>Shortest path</td>
<td>0.013</td>
<td>0.071</td>
<td>0.214</td>
<td>0.208</td>
</tr>
<tr>
<td>Edges</td>
<td>0.253</td>
<td>0.078</td>
<td>0.362</td>
<td>0.369</td>
</tr>
<tr>
<td>Nodes</td>
<td>0.022</td>
<td>0.067</td>
<td>0.368</td>
<td>0.378</td>
</tr>
<tr>
<td>Transitivity</td>
<td>0.010</td>
<td>0.083</td>
<td>0.014</td>
<td>0.005</td>
</tr>
</tbody>
</table>

\[ p_{value} > 0.05 \quad p_{value} < 0.05 \]
## Time series

### ARIMA fittings

<table>
<thead>
<tr>
<th>Network metric</th>
<th>Value of $d$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Clustering</td>
<td>55</td>
</tr>
<tr>
<td>Betweenness centrality</td>
<td>57</td>
</tr>
<tr>
<td>Cliques</td>
<td>69</td>
</tr>
<tr>
<td>Diameter</td>
<td>60</td>
</tr>
<tr>
<td>Intermittency</td>
<td>56</td>
</tr>
<tr>
<td>Load centrality</td>
<td>63</td>
</tr>
<tr>
<td>Degree</td>
<td>51</td>
</tr>
<tr>
<td>Radius</td>
<td>58</td>
</tr>
<tr>
<td>Shortest path</td>
<td>55</td>
</tr>
<tr>
<td>Edges</td>
<td>61</td>
</tr>
<tr>
<td>Nodes</td>
<td>49</td>
</tr>
<tr>
<td>Transitivity</td>
<td>64</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>698</strong></td>
</tr>
</tbody>
</table>

**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 $p + d + q$.

<table>
<thead>
<tr>
<th>Book</th>
<th>Measure</th>
<th>Sum</th>
<th>p</th>
<th>d</th>
<th>q</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Poems of Jonathan Swift, D.D., Volume 2</td>
<td>Load centrality</td>
<td>9</td>
<td>5</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>The Journal to Stella</td>
<td>Clustering</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>The Iron Heel</td>
<td>Clustering</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Typee: A Romance of the South Seas</td>
<td>Edges</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
Data analysis

Dimensionality reduction
- Feature selection
- Feature extraction

Supervised learning
- Zero Rule: \( \frac{1}{8} = 12.5\% \)
- One Rule
- Naive Bayes
- K-Nearest Neighbors
- J48 (tree)
- Radial Basis Function Networks

Precision
\[
P_A = \frac{TP_A}{TP_A + FP_A}
\]

Recall
\[
R_A = \frac{TP_A}{TP_A + FN_A}
\]

48 Attributes
80 Books
8 Authors

\(TP\): True Positives \(FP\): False Positives
\(FN\): False Negatives

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

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.

- Arthur Conan Doyle
- Bernard Shaw
- Fyodor Dostoyevsky
- Herman Melville
- Jack London
- Jonathan Swift
- Leo Tolstoy
- Nathaniel Hawthorne
Feature extraction

Scores using feature extraction.
Summary of classification success scores

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