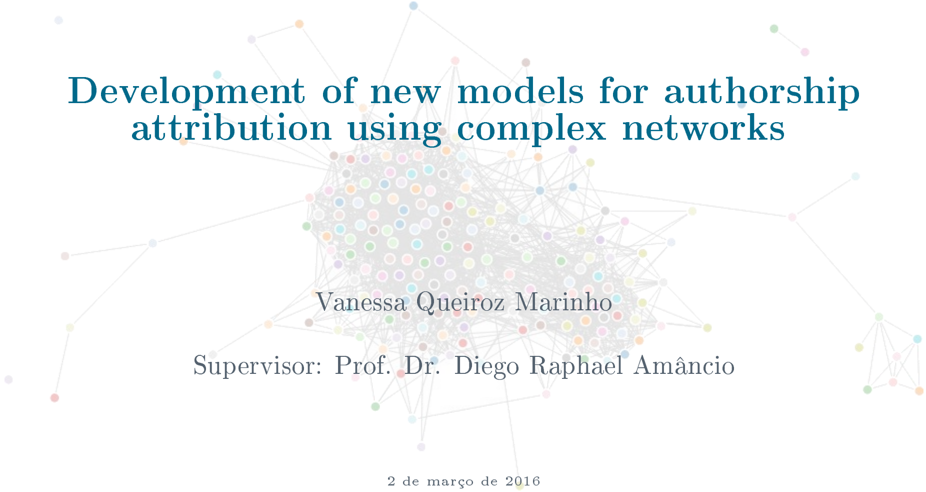
Institute of Mathematical and Computer Sciences
University of São Paulo



Development of new models for authorship attribution using complex networks

Vanessa Queiroz Marinho

Supervisor: Prof. Dr. Diego Raphael Amâncio

2 de março de 2016

Agenda

Introduction

Background

- Complex Networks
- Authorship Attribution

Work Proposal

- Motivation and Goals
- Methods
- Preliminary Results
- Ongoing Works

Introduction

The modelling of **real systems** using **complex networks** is useful to describe a variety of systems [1].

The **textual networks** are important to the development of this project. In special, the networks created by **syntactical relations**.

A particular case of syntactical networks are the **co-occurrence networks** (or networks of adjacency of words). We aim to use this discrimative power to characterize writing styles in the authorship attribution problem.

[1] Albert, R. and Barabási, A.-l. *Statistical mechanics of complex networks*. Rev. Mod. Phys, 2002.

Agenda

Introduction

Background

- Complex Networks
- Authorship Attribution

Work Proposal

- Motivation and Goals
- Methods
- Preliminary Results
- Ongoing Works

Complex Networks

Definition

Complex Networks have been used as the mathematical representation of a variety of complex systems.

The study of networks was limited to graph theory applied to various **random** systems. One of the precursors of graph theory was the mathematician Leonhard Euler.

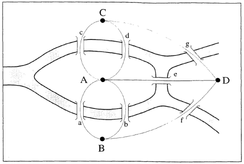


Figura 1: Kőnigsberg bridges. Figure extracted from [2]

A network $G = \{V, E\}$ is formed by a set $V = \{v_1, v_2, ..., v_n\}$ with nodes and another set $E = \{e_1, e_2, ..., e_m\}$ with edges.

[2] Barabasi, A.-L.*Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life.* Plume Books, 2003.

Complex Networks

Network Models

According to Newman [3], the mathematical models allow to **understand the effects of different properties** on the networks. The main models are cited below:

- Erdős-Rényi (ER) Model [4]
- Watts-Strogatz (WS) Model or Small-World networks [5]
- Barabási-Albert (BA) Model or Scale-Free networks [6]

[3] Newman, M. *Networks: An Introduction*. Oxford University Press, 2010

[4] Erdös, P. and Rényi, A. *On Random Graphs I*. Publicationes Mathematicae Debrecen, 1959

[5] Watts, D.J. and Strogatz, S.H. *Collective dynamics of 'small-world' networks*. Nature, 1998.

[6] Barabasi, A.-L. and Albert, R. *Emergence of Scaling in Random Networks*. Science, 1999.

Erdős-Rényi (ER) Model: [7]

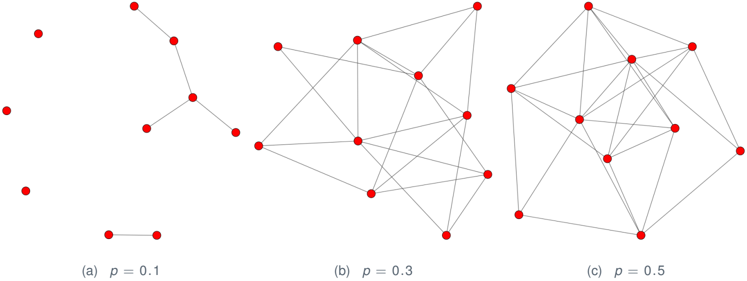


Figura 2: ER Model with 10 nodes and different *p* probabilities of connecting edges.

Problems: A few cycles and the degree distribution isn't a power law.

[7] Erdös, P. and Rényi, A. *On Random Graphs I*. Publicationes Mathematicae Debrecen, 1959

Watts-Strogatz (WS) Model or Small-World networks: [8]

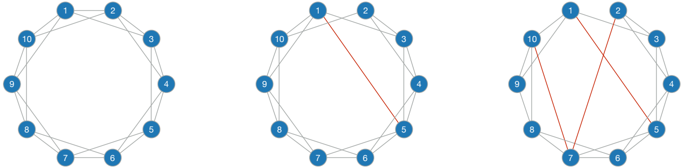


Figura 3: Reconnecting process on WS networks

Characteristics: High clustering coefficient and short distances.
Problems: Distribution is not a power law.

[8] Watts, D.J. and Strogatz, S.H. *Collective dynamics of 'small-world' networks.* Nature, 1998.

Barabási-Albert (BA) Model or Scale-Free networks: [9]

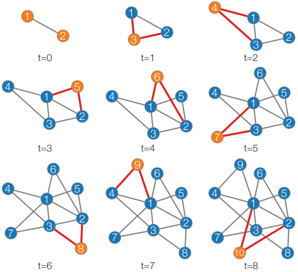


Figura 4: Adding new nodes at the BA model

Characteristics: High clustering coefficient, degree distribution is a power law, *hubs*.

[9] Barabasi, A.-L. and Albert, R. *Emergence of Scaling in Random Networks*. Science, 1999.

Complex Networks

Networks used to analyse the language

Complex networks are used to model and analyse the human language [10] which is also considered a **complex system**.

$N = (V, E)$, where V is the set of vertices representing the **linguistic units** and E is the set of edges representing the **relations** between these units.

Linguistic units:

- Words
- Phoneme
- Morpheme

Relations:

- Co-occurrence
- Syntactical
- Semantical

[10] Cong, J. and Liu, H.*Approaching human language with complex networks.* Physics of life reviews, 2014.

Complex Networks

Dorogovtsev-Mendes(DM) Model

Dorogovtsev and Mendes proposed a theory about the **evoluation of the human language**:

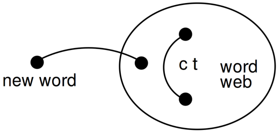


Figura 5: The growth of the word network. Figure extracted from [11]

A new word n is added to the network (total of nodes is equal to t).

1. n is connected with a word i with probability proporcional to k_i
2. ct new edges are added between words i and j with probability proporcional to $k_i k_j$

[11] Dorogovtsev, S. N. and Mendes, J. F. F. *Language as an evolving word web*. Proceedings of the Royal Society of London, 2001

Complex Networks

Complex Networks measurements

The main measurements used in this project are below:

- Degree
- Assortativity
- Average degree of the neighbours
- Clustering Coefficient
- Average of the shortest paths
- Betweenness
- Accessibility

Agenda

Introduction

Background

- Complex Networks
- Authorship Attribution

Work Proposal

- Motivation and Goals
- Methods
- Preliminary Results
- Ongoing Works

Tipical problem: a text of unknown authorship is attributed to an author from a set of possible authors.

Mosteller e Wallace started the studies based on **statistical methods**. They analysed the authorship of different political essays, known as *The Federalist Papers*.

Authorship Attribution

Stylometric features

Many works define characteristics that quantify the writing style of an author, known as **stylometry**.

The stylometric features are devided in the following categories [12]:

- Lexical features
- Character-based features
- Syntactical features
- Semantical features

[12] Stamatatos, E. *A Survey of Modern Authorship Attribution Methods*. J. Am. Soc. Inf. Sci. Technol.,2009.

Authorship Attribution

Attribution Methods

Profile-based Methods

Instance-based Methods

Agenda

Introduction

Background
- Complex Networks
- Authorship Attribution

Work Proposal
- Motivation and Goals
- Methods
- Preliminary Results
- Ongoing Works

Work Proposal

Motivation

The huge amount of text available on the Web releaved the **potential of authorship analysis** in different applications.

This task is relevant inside the natural language processing area contributing to advances in [13]:

- literature
- history
- inteligence services
- forensics
- criminal investigations
- plagiarism

[13] Stamatatos, E. *A Survey of Modern Authorship Attribution Methods*. J. Am. Soc. Inf. Sci. Technol.,2009.

Work Proposal

Goals

The main goals of this work are:

- Develop **new models**, including adapting the co-occorrence model.
- Introduce **new measurements**.
- Combine **topological and traditional attributes** in hybrid classifiers

Agenda

Introduction

Background

- Complex Networks
- Authorship Attribution

Work Proposal

- Motivation and Goals
- Methods
- Preliminary Results
- Ongoing Works

Methods

Database

40 books, 5 of each author, published between 1835 and 1922.

- Arthur Conan Doyle
- Bram Stoker
- Charles Dickens
- Edgar Allan Poe
- Hector Hugh Munro (Saki)
- Pelham Grenville Wodehouse
- Thomas Hardy
- William Makepeace Thackeray

Methods

Pre-processing

Removing *stopwords* and lemmatization.

Original Text	No *stopwords*	After lemmatization
“There are three men waiting for	*three men waiting*	*three men wait*
him at the door", said Holmes.	*door said holmes*	*door say holmes*
“Oh, indeed! You seem to have	*oh indeed seem*	*oh indeed seem*
done the thing very completely.	*done thing completely*	*do thing completely*
I must compliment you."	*must compliment*	*must compliment*
“And I you", Holmes answered.	*holmes answered*	*holmes answer*

Methods

Connecting words

Networks are **undirected** and **unweighted**.

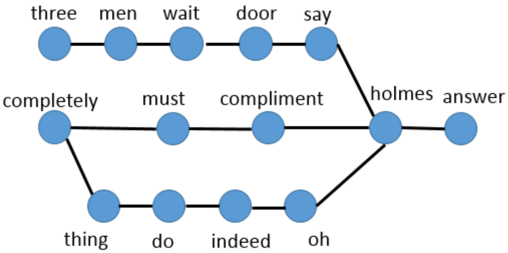


Figura 6: Graph that represents the traditional co-occurrence network for the sentence "*three men wait door say holmes oh indeed seem do thing completely must compliment holmes answer*

”

Methods

Extracting global properties from local properties

Almost all measurements are extracted from each one of the words.

Average:
$$\langle X \rangle = \frac{1}{M} \sum_{i=1}^{M} X_i \tag{1}$$

Deviation:
$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{M} (X_i - \langle X \rangle)^2}{M - 1}} \tag{2}$$

Skewness:
$$\gamma(X) = \left\langle \left(\frac{X - \langle X \rangle}{\sigma(X)} \right)^3 \right\rangle \tag{3}$$

Agenda

Introduction

Background

- Complex Networks
- Authorship Attribution

Work Proposal

- Motivation and Goals
- Methods
- Preliminary Results
- Ongoing Works

Preliminary Results

Modelling extension

The co-occurrence model does not capture possible relations between distant words. Some alternatives:

- *Further Neighborhoods*: All word pairs separated by at most $W - 1$ words are connected, where $W = 1, 2, 3$.
- Connecting all words inside a sentence, approach called *Sentence based*.

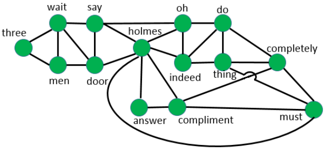


Figura 7: Graph that represents the approach *Further Neighborhoods* with $W = 2$ for the sentence *three men wait door say holmes oh ...*

Preliminary Results

Modelling extension

Modelling	C4.5		kNN		SVM		Naive Bayes	
	AA	FS	AA	FS	AA	FS	AA	FS
$W = 1$	27.5%	55.0%	50.0%	62.5%	52.5%	55.0%	47.5%	62.5%
$W = 2$	45.0%	67.5%	62.5%	72.5%	55.0%	60.0%	50.0%	57.5%
$W = 3$	45.0%	57.5%	55.0%	60.0%	55.0%	62.5%	50.0%	65.0%
Sentence based	32.5%	47.5%	42.5%	65.0%	55.0%	62.5%	40.0%	55.0%

- The *Further Neighborhoods* approaches with $W = 2$ and $W = 3$ shown better results. Which **confirms the initial hypothesis** that the connection of words in a 'bigger' context improves the authorship attribution performance.

Preliminary Results

Motifs extraction

The extracted motifs are:

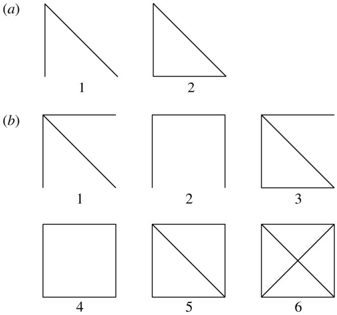


Figura 8: All undirected motifs with 3 nodes (a) and 4 nodes (b). Figure adapted from [15]

[14] Silva, E. and Stumpf, M. P. H. *Complex networks and simple models in biology*. Journal of the Royal Society Interface, 2005.

[15] Silva, E. and Stumpf, M. P. H. *Complex networks and simple models in biology*. Journal of the Royal Society Interface, 2005.

Preliminary Results

Motifs extraction

The extracted motifs are:

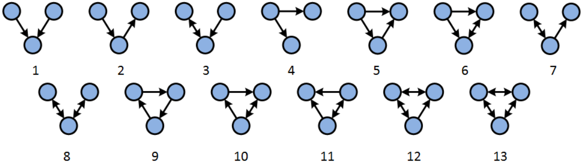


Figura 9: All directed motifs with 3 nodes. Extracted from [16]

[16] Gabasova, E. *Analysing programming languages using dependency networks*, 2014. Available at `http://evelinag.com/blog/2014/06-09-comparing-dependency-networks/index.html`.

Preliminary Results

Motifs extraction

Modelling	**C4.5**	**kNN**	**SVM**	**Naive Bayes**
$W = 1$	30.0%	32.5%	27.5%	27.5%
$W = 2$	55.0%	40.0%	35.0%	30.0%
$W = 3$	42.5%	35.0%	32.5%	32.5%

Tabela 1: Percentage of books correctly classified using the 13 directed motifs

Modelling	**C4.5**	**kNN**	**SVM**	**Naive Bayes**
$W = 1$	30.0%	35.0%	27.5%	35.0%
$W = 2$	42.5%	40.0%	37.5%	42.5%
$W = 3$	40.0%	40.0%	40.0%	45.0%

Tabela 2: Percentage of books correctly classified using the 8 undirected motifs

Agenda

Introduction

Background

- Complex Networks
- Authorship Attribution

Work Proposal

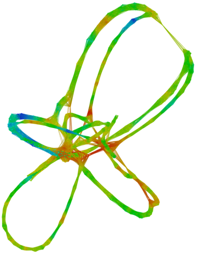
- Motivation and Goals
- Methods
- Preliminary Results
- Ongoing Works

Ongoing Works

Besides the discribed activities, there are some ongoing works with researches from the Physics department

- Analysis of Twitter data.
- Development of new textual modelings.

Figura 10: Network of paragraphs similarity from the book *The Adventures of Sally*. The accuracy rate of the authorship attribution task reached 62.5%, with the usage of simetry measurements.



Thank you!

Acknowledgements:

CAPES FAPESP USP