Generating co-evolutionary polarized opinion networks

Victor Morel
Abstract

Generating co-evolutionary polarized opinion networks

Victor Morel

In this thesis we present a co-evolutionary opinion network model, which aim to efficiently represent an online social network, with each person represented as a node attributed with an opinion, and the relations between people by edges. For this, we needed to classify the existing models through the prism of co-evolution, i.e. how the topology and the state of the network interact between each other, so one can find a model of classification for network models. This model will be compared under certain aspects to empirical data, as well as previous works. We show the emergence of a polarization of the opinions in the network, which appears only under certain conditions: a strong homophily between nodes, as well as a co-evolutionary behavior, i.e. a strong interaction between the topology and the state, of the model.
“Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.”

George E. P. Box

“If you torture the data long enough, it will confess to anything.”

Darrell Huff (How to Lie With Statistics, 1954)
Acknowledgements

I’d like to thank my supervisor Stéphane BONNEVAY, as well as my reviewer Matteo MAGNANI, but also my family and friends, without whom I would never have acquired the motivation to go abroad and accomplish what I did, so little it would be.
Contents

Acknowledgements v

1 Introduction 1
1.1 Text-Networks 1
  1.1.1 Opinion representation 1
  1.1.2 What is a network? 2
  1.1.3 And a social one? 2
  1.1.4 What we exchange online 3
1.2 Co-evolution 3
  1.2.1 What is the relation between the text and the structure in a social network? 4
1.3 Polarization 4
1.4 Why did we choose to study it? 4
  1.4.1 A new co-evolutionary model ... 5
  1.4.2 ... And a new co-evolutionary classification 6
1.5 Our approach 6

2 Related work 7
2.1 Co-evolutionary classification 7
  2.1.1 Adding a dimension 7
  2.1.2 Basic Models 10
     State 10
     Topology 10
  2.1.3 TS Models 12
     TS Classic 12
     TAS 13
     TSA 13
     TASA 13
  2.1.4 ST models 14
     ST Classic 14
     STA 15
     SAT 15
     SATA 15
  2.1.5 Co-evolutionary 16
     STS Classic 17
     STAS 17
     SATS 18
     SATAS 18
  2.1.6 Details of highly relevant papers 18
     DosSantos 18
     Lee 19
3 The model

3.1 Construction ........................................... 21
  3.1.1 Why a model? .................................... 21
  3.1.2 Tools ............................................... 21

3.2 Our implementation .................................... 22
  3.2.1 Core ............................................... 22
  3.2.2 Choice of node Product ................................ 22
      Sum .................................................. 23
  3.2.3 Summary Table .................................... 24
  3.2.4 State update ...................................... 26

3.3 Explanation of the parameters .......................... 26
  3.3.1 Of the importance of initialization ................. 26
      Different starts .................................... 26
      Choice of attributes ................................ 27
  3.3.2 Topology update ................................... 27
      Number of edges and nodes .......................... 27
      Tuning $\sigma$ ...................................... 27
      Tuning $\theta$ ....................................... 27
  3.3.3 State update ...................................... 27
      Speed of convergence ................................ 28
      Weight of convergence ............................... 28

3.4 Hypothesis ............................................ 28

4 Experiments .................................................. 29
  4.1 Datasets ............................................... 29
      4.1.1 Political blog ................................... 29
  4.2 Analysis Tools ......................................... 29
      4.2.1 Average Attribute Deviation ...................... 29
      4.2.2 Bimodality ...................................... 29
      4.2.3 Power-law Coefficient ........................... 31
      4.2.4 Polarization .................................... 32
      4.2.5 Others .......................................... 32
  4.3 Results of the experiments ............................. 32
      4.3.1 Hypothesis 1 .................................... 32
      4.3.2 Hypothesis 2 .................................... 38
      4.3.3 Hypothesis 3 .................................... 39
      4.3.4 Hypothesis 4 .................................... 41
  4.4 Analysis ................................................ 41

5 Discussion and further work ................................ 43
  5.1 Conclusion ............................................. 43
  5.2 Limitations ............................................ 43
      5.2.1 Conceptual limitations .......................... 43
      5.2.2 Technical limitations ............................ 43
  5.3 Possible enhancements .................................. 43
      5.3.1 Hypothesis 5 .................................... 43
      5.3.2 Measure of polarization .......................... 44
      5.3.3 Working with text ................................ 44
      Topic Models ......................................... 44
      Natural Language Processing ......................... 45
A Comparison: product vs sum

A.1 Peculiar case 1

A.1.1 Description

A.1.2 Product behavior

A.1.3 Sum behavior

A.2 Peculiar case 2

A.2.1 Description

A.2.2 Product behavior
  Low $\sigma$
  High $\sigma$

A.2.3 Sum behavior

A.3 Peculiar case 3

A.3.1 Product behavior

A.3.2 Sum behavior

A.4 Equilibrium between probabilities

A.4.1 Product

A.4.2 Sum

Bibliography
List of Tables

2.1 Overview of non-trivial network models through the prism of co-evolution. .......................... 9
2.2 Details of the Basic Models .......................... 11
2.3 Details of the TS Models .......................... 12
2.4 Details of the ST Models .......................... 14
2.5 Details of the STS Models, also known as co-evolutionary .......................... 16
3.1 Comparison product and sum formulas : summary table .......................... 25
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNA</td>
<td>Social Network Analysis</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>TS</td>
<td>Topology (to) State</td>
</tr>
<tr>
<td>ST</td>
<td>State (to) Topology</td>
</tr>
<tr>
<td>STS</td>
<td>State (to) Topology (to) State</td>
</tr>
<tr>
<td>AAD</td>
<td>Average Attribute Deviation</td>
</tr>
<tr>
<td>BC</td>
<td>Bimodality Coefficient</td>
</tr>
<tr>
<td>KDE</td>
<td>Kernel Density Estimation</td>
</tr>
<tr>
<td>BA</td>
<td>Bimodality Amplitude</td>
</tr>
<tr>
<td>BI</td>
<td>Bimodality Index</td>
</tr>
</tbody>
</table>
Pour ma grand-mère …
Chapter 1

Introduction

1.1 Text-Networks

1.1.1 Opinion representation

Nowadays, 3.2 billion people use internet (United Nations 2015), and among these, around three quarters (Pew research center) are connected to social media websites, which gives us not less than 2.4 billion people connected to Twitter, Facebook, Pinterest etc. On these platforms, we communicate via natural language, whether English, Indian, Chinese, and these messages mean something: they express what we think, as limited as they can be.

Often the messages, posts and retweets are aimed to represent our opinion on a topic, therefore with the help of network science, the area of this thesis, we can have a big picture of the main tendencies.

For example, in figure 1.1, we can visually see that in the US political blogosphere in 2004, there are two clear tendencies: liberal and conservative, and these two entities are very connected within, but sparsely with the other.

![Figure 1.1: Political Blogs in the US in 2004, data compiled by Lada Adamic](image)

This example is quite clear, but this is not always the case, and most of the time we need mathematical tools to represent such networks, to try to recreate them and see if we understand the underlying mechanisms behind
Chapter 1. Introduction

the dynamics: these are called models, and for this work we needed a peculiar one, so we studied and compared the existing, to realize that we had to build and test a new one.

1.1.2 What is a network?

In nature, numerous entities can be represented as networks, such as biological, physical or social phenomena.

Networks are usually represented as graphs, where the agents of each network are represented as nodes (or vertexes), and the links between them as edges. The pattern of connections forms what is called the topology of the network (also known as dynamics of networks).

The study of network formation is not new (Euler in 1736, Sylvester in 1878, Moreno in 1933 etc), and it has been realized that most real life networks have properties, features that are not found if you generate a graph randomly.

It was also figured that local simple laws give complex properties to the big systems, so we have been trying to conceive models to artificially recreate networks with real world properties, to study them. Networks exhibiting such features are called complex networks.

In order to generate network models, one needs graphs constructed randomly: a network model in which some specific set of parameters take fixed values, but where the network is random in other aspects. Some classify these models between Edge Assignment (the size of the network is fixed at creation, and edges are added or removed according to particular dynamics) and Network Growth (the network gains nodes as well as edges), such as the work of Dickison, Magnani, and Rossi [1], but for this project we will need a different approach of classification that we will discuss later.

1.1.3 And a social one?

As mentioned, the main focus of this work will be on social networks, but what is a social network, and what are the main features, does it differ from other kinds of networks?

We can define a social network as a social structure made of agents (people, organizations etc), and the relationship between these agents. What we study through Social Network Analysis is only the representation of such a structure in terms of graphs and network science. This discipline has existed for more than a century now, as we can trace its roots to the work of Emile Durkheim, the father of sociology, and has been an active field for 70 years now.

Why are we talking so much about it nowadays then? Because of the rise of Social Media, and all the big data that comes with it. Actually, even with just numeric data, we started to have something to work on. As we
started to develop tools to study all kind of networks, we also started to 
grasp that these social networks were different from the other types of net-
works such as biological networks or technological networks (see Newman
and Park [2]).

Social networks are found to be degree-assortative, which means that a
node with a large number of connections (a high degree), will tend to be
connected to other nodes with the same number of connections, and con-
versely. Another aspect of social networks that we do not naturally observe
in other types of networks is the presence of community structures: usu-
ally we refer a community as a part of a network where the connections are
very dense inside, and sparse with the outside.

1.1.4 What we exchange online
If you are familiar with Twitter, if you have a Facebook account, or use any
widely spread social media, you already know what a text-network is, but
you might not realize that you know. A text-network is a way to represent
a social network, by associating text to each agent, quite simple right? In
appearances, yes, but in practice the analysis of such a representation of
social networks is extremely difficult, because one needs on one hand net-
work skills, network science still being an emerging field, and text-mining
skills on the other hand.

The text in text-network can mean various things, such as the messages
we exchange, but more particularly the size of the message, the topic we
are talking about, the register we employ etc.

But these attributes, those features each person owns, are difficult to
manipulate and complex to analyze. A simpler way to represent them is
as opinions, our view on a topic. This representation has drawbacks, not-
ably because it is simplifying: the representation is limited to a topic, or
a set of topic unlike text in its global meaning which can be interpreted,
parsed and dissected in many different ways. But it also has advantages:
we can represent the state of a network according to a particular subject (or
like a general tendency) as a binary state (pro or against), and without any
difficulties as something more nuanced (between completely pro and com-
pletely against). We are aware that an opinion cannot always be reduced
as a simple yes or no on a question, but for the purpose of the work we
decided to go with this representation.

1.2 Co-evolution
To understand how a network works, a way to do so is to study how it
grows. And to understand how it grows, one needs to understand what are
the factors that conditioned its evolution. The first and main focus of study
is the topology itself: de Solla Price [3] imagined the first model where
a node will choose its new connection according to the current number of
connections of existing nodes.

But the state of a node can as well influence a network’s evolution: the
SIR epidemic model (first imagined by Kermack and McKendrick [4]) takes into consideration the neighbors’ state of a node to determine if the latter will get infected, or will recover from a disease. But as the influence of the topology has been largely studied, the influence of the state of individual agents inside a network is still an open field.

1.2.1 What is the relation between the text and the structure in a social network?

Among the different representations of the state, one that is entirely virgin to research is the text contained in a network: it could be what we think (and post as a status), the messages we exchange, the content of what we share etc. How much does the text influence the evolution of a network, under what circumstances? Is it the size of the exchanged messages, the subject of a conversation, the register used (in a sociolinguistics meaning), maybe the frequency with which hashtags are used...? And conversely: does the network itself (as a complex system) influence the aforementioned text? Most likely, but to what extent? All those questions lead us to a more global one: what is the relation between the attributes and the structure of a social network in the latter? To what extent do they influence each other?

1.3 Polarization

Sometimes, in certain social networks, we can observe a division between two highly cohesive communities: it is referred as polarization (and it has to be understood as social polarization).

A certain number of works studied this phenomenon from a sociological point of view, but also from a computer scientist angle, but none of them had a focus on the creation from a network scientist perspective: what are the interactions, the underlying mechanisms that make a network polarized? Webster [5] and more recently Lynch [6] hypothesized that social media could potentially play a role in encouraging social polarization because they help creating homophilous communities.

In this thesis, we explore the emergence of polarization, and if the simple argument of homophily in a network is sufficient to create polarization, or if another characteristic is necessary or not. We therefore investigated this new track which is coevolution in a network, combined with affinity, and present our results in a following section (chapter 4).

1.4 Why did we choose to study it?

The idea itself is not new, even if quite recent. Indeed, network science and text-mining in social media are both emerging fields, but only recently that we actually have had the tools to combine the two areas. Without this notion of attributed network, the scientific community could
1.4. Why did we choose to study it?

think only in terms of topology (cf preferential attachment), but now that we have the tools and the expertise required to model social networks in terms of attributed graphs, we can extrapolate this concept of attributes to opinion.

This change not only opens us a broad new range of studies, it also grants us a better understanding of each agent as an entity within a network (because we can model each person not only as a bare node, but now as something more advanced).

Scientific papers efficiently considering attributed networks, and homophilic networks, have been published in 2015 (see [7] and [8]), this seems like a logical consequence but to pursue in that direction now that it is possible, now that our understanding of the prerequisites are fulfilled.

1.4.1 A new co-evolutionary model ...

We introduced earlier the topology of a network as the dynamic OF networks, but with the emergence of another major line of research about the state of the network, also known as dynamics ON networks, where each node represents a dynamical system, we started to study not only one of the dynamic, but both at the same time : how they are correlated (in a generic meaning).

To fulfill that purpose, existing network models were insufficient : they are either too focused on the topology, or on the state, whereas what we needed here is a network model which studies the mutual influence of both. A model that exhibits both dynamics is called a co-evolutionary model: it has a feedback loop between the state and the topology, see figure 1.2

Figure 1.2: Co-evolutionary dynamics, from Gross and Blasius [9]

Thinking in terms of co-evolutionary networks allows us to study the relation between the structure and what’s inside each node when the network is evolving. But even the few recent models including both dynamics are not fit for SNA, and especially the study of text, so we had to create another one, more complete, and more suited to this project.

The creation of a model with a co-evolutionary behavior for social networks has never been attempted before, so we expect the implementation of this new part to be hazardous.
1.4.2 ... And a new co-evolutionary classification

As we were reviewing the different kind of network models, we became aware of the lack of overview surrounding the topic (notably because of the novelty of the preceding) ; we needed a tool to comprehend our own needs, and why the existing models were inadequate for our issue. Therefore, the following section is an overview of network models and applications, through the prism of co-evolution.

1.5 Our approach

At first we thought about approaching the problem by mining some empirical data, then extracting patterns from that data, but we needed first to build a model able to apprehend the functioning of opinion networks. We have based our work on existing models, that we compared and enhanced, this part is detailed in chapter 2, where we provide the keys for understanding, and a first version of the classification. The model in itself is presented in chapter 3, with a theoretical analysis, the details of each parameter and the influence of the latter on the results. Our experiments are exposed in chapter 4, they show that our model provides a good fit of many natural graph statistics. Finally the conclusion of this thesis can be found in the chapter 5, a critical analysis of its limitations, and what could be possibly done to improve it.
Chapter 2

Related work

2.1 Co-evolutionary classification

The main area for this work is network science, and more precisely co-evolutionary networks. As aforementioned, this field is varied and heterogeneous, and we observed that it was sometimes hard, occasionally impossible to compare and sort network models, and thus to highlight the useful differences between the latter.

We decide to present the related work classified according to their relation with co-evolution, to have a better idea of what they are doing, how they behave, and to navigate more easily between each model. This classification is a contribution to this thesis. It is to be noted that for the short duration of a master thesis, it was extremely hard to find an example or a reference for each and every type of model, for that reason some cases are empty.

2.1.1 Adding a dimension

We provide an explanation of the new tools we are working on:

- The **Topology** of a network is the set of edges and vertexes.

- The **State** is the set of attributes attached to the set of nodes.

$\rightarrow$ **Entity** An incoming arrow means that the entity is influenced by the other one, therefore is not static.

$\leftarrow$ **Entity** An outgoing arrow means that the entity influences the other one, without giving information on the dynamicity of the latter.

The two previous arrows don’t give information on the temporal evolution of the model: it can grow, or not, that axis is decided by:

- A loop means that the entity affects itself, for that purpose we have to take into account the time: it can only affects itself within two different time-steps

---

1It has to be noted that in the case where there is only two time steps: the original and the final one, we consider it as non-temporal
Chapter 2. Related work

A loop gives informations on the temporal evolution of a model according to a flow representation, therefore if there is a loop on top of an entity, and this entity affects the other, then the second entity will evolve according to time.

If the first entity doesn’t have any loop, then it’s considered as static.

Considering the naming strategy:

A TS model is characterized by the influence of the Topology on the State

A ST model is characterized by the influence of the State on the Topology

A STS model is characterized by a feedback between the two entities, it’s also known as co-evolutionary or adaptive network.

If a $A$ is following a $S$ or a $T$, it means that the model is Aware of its own State or Topology (respectively), and that it influences itself. It is equivalent to the loop in the symbols.

Now an overview is provided as a framework of interpretation.

\footnote{The case where an entity is updated randomly (neither from an intern nor an extern influence) is not taken into account. We consider this case as non-relevant for study, especially in this field.}
### Table 2.1: Overview of non-trivial network models through the prism of co-evolution.

<table>
<thead>
<tr>
<th>TS Models</th>
<th>ST Models</th>
<th>STS Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Diagram](TS Models)</td>
<td>![Diagram](ST Models)</td>
<td>![Diagram](STS Models)</td>
</tr>
</tbody>
</table>

Models whose main characteristic is the influence of the Topology on the State

Models whose main characteristic is the influence of the State on the Topology

Co-evolutionary models
Chapter 2. Related work

2.1.2 Basic Models

These models are not necessarily basic, but they can be considered as such from a co-evolutionary point of view.

State

This aspect was reserved for the physicists for a long time, because the study of dynamics on networks was thought to be more appropriate in this field. But nowadays situation is changing, and the physicists don’t have the monopoly, notably because of the development of the epidemics models (those without considering the topology, such as Kermack and McKendrick [4]), which are relevant to be studied also from an computer scientist point of view. It’s highly prevalent that there is a correlation with the topology, but we can find some examples where there isn’t.

We can classify random variables, Markov processes, as well as some basic epidemic models in this genus.

Topology

These models enable the study of structure of the network, without considering the intern dynamics.

Edge Assignment  You start with a fixed number of nodes, and add edges, like Erdős and Rényi [10]. DD is poisson, but no correlations between the degrees, nor a high CC.

Small World  It’s also an EA model, but without radically different features (low diameter) Watts and Strogatz [11].

Network Growth  In the case of Social Network Analysis, the Network Growth class of models is more appropriate

You start with one node, or a small number of nodes, and you add nodes and edges
### 2.1. Co-evolutionary classification

#### Basic Models

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Applications and Models</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="State" /></td>
<td>Random Variables</td>
<td></td>
</tr>
<tr>
<td><img src="image2.png" alt="State" /></td>
<td>SI(R) (non-)Markov processes</td>
<td>[4]</td>
</tr>
<tr>
<td><img src="image3.png" alt="Topology" /></td>
<td>E-R Random Graph, Small-World model</td>
<td>[10] [11]</td>
</tr>
<tr>
<td><img src="image4.png" alt="Topology" /></td>
<td>Preferential attachment and derivative works</td>
<td>[12]</td>
</tr>
<tr>
<td><img src="image5.png" alt="State" /> <img src="image6.png" alt="Topology" /></td>
<td>Dynamical system in a network</td>
<td></td>
</tr>
<tr>
<td><img src="image7.png" alt="State" /> <img src="image8.png" alt="Topology" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image9.png" alt="State" /> <img src="image10.png" alt="Topology" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 2.2: Details of the Basic Models**
### 2.1.3 TS Models

Now there is at least one influence between the State and the Topology. Basically epidemics, gossip, and PageRank as well

"The modeling of infectious diseases is a tool which has been used to study the mechanisms by which diseases spread, to predict the future course of an outbreak and to evaluate strategies to control an epidemic" \(^3\). It appears that these models are mainly concerned by the effect of the topology’s evolution on the state of the networks.

#### TS Classic

\(^3\)Attributed to Daley & Gani, 2005
2.1. Co-evolutionary classification

You can find various examples such as SI, SIR, SIRS etc in Newman [16]. The main (and simpler) model is the SI model, where $S$ stands for Susceptible and $I$ for Infected. An agent can be in either one of this state, and a node in a $S$ state linked to (at least) a $I$ has a probability $\beta$ to be infected. The SIR model lets the possibility for an agent to recover with a probability $\gamma$ (state $R$ for recover), in the SIS model an agent can be infected again etc. As the speed of infection depends on degree (see Gross and Sayama [17]), there is clearly an influence of the topology on the state.

The closest work from our own is Morales, Borondo, Losada, et al. [14]: they present the diffusion of information on Twitter, more precisely a methodology to study and measure the emergence of polarization from social interactions. They also propose a model to estimate opinions in which a minority of influential individuals propagate their opinions through a social network. Their last but not least contribution is the introduction of a polarization index, to measure how much a network is polarized. The latter, although not as reliable as announced, was a solid base for the evaluation of our project.

Some models are more evolved, such as Qin, Zhong, Jiang, et al. [15], where they present an epidemic model with self-awareness of the environment. Here the agents have a strategy to avoid the spreading of the infection: they have a lower probability to connect to an infected node (therefore, there is reflexivity on the state).
Chapter 2. Related work

<table>
<thead>
<tr>
<th>ST Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
</tr>
<tr>
<td><img src="State" alt="State" /> <img src="Topology" alt="Topology" /></td>
</tr>
<tr>
<td><img src="State" alt="State" /> <img src="Topology" alt="Topology" /></td>
</tr>
<tr>
<td><img src="State" alt="State" /> <img src="Topology" alt="Topology" /></td>
</tr>
<tr>
<td><img src="State" alt="State" /> <img src="Topology" alt="Topology" /></td>
</tr>
</tbody>
</table>

**Table 2.4: Details of the ST Models**

### 2.1.4 ST models

Computer scientists have been trying to build models with a reciprocal effect between the state and the topology of the network (dynamics on networks vs dynamics of networks), but it’s still an emerging field, and the current models don’t fully capture the influence of state on topology. Usually classified as Homophilic Networks.

**ST Classic**

Definition of graphs with attributes

In 2010, Kim and Leskovec [18] started to formalize the concept of graphs with attributes, but without the preferential attachment. In brief, it studies the effect of the state on the topology, but doesn’t take into account the
topology to update the state. This MAG model enables a very detailed analysis of the statistical properties of a graph (and therefore a network).

**Generalized Scale-free Homophilic**  Dos Santos, De Almeida, Mendes, *et al*. [7] present a model where each node has an attribute, and this attribute (to be more precise, the similarity between each node’s attribute) will impact the topology, as well as the topology will influence itself. Indeed, unlike the Kim and Leskovec [18] model, Dos Santos, De Almeida, Mendes, *et al*. [7] use the degree to determine new links (as in Barabási and Albert [12]), the state affects the topology, and the topology affects itself, but the impact of the topology on the state is still missing.

**Preferential Attachment in Graphs with Affinities**  Lee, Stephan, and Smola [8] built a model with an vector of attributes for each node, and the evolution of the model includes the similarity between the nodes.

Based on Buckley and Osthus [19] for the power-law DD, and Blei and Frazier [20] for the affinity between nodes.

But if we go in the details, indeed the state of the nodes participates to the determination of the topology, but the topology doesn’t really influence the state, the modification of the pattern of connection doesn’t change the nodes’ attributes, it just helps the model to be more precise (with a weight matrix and logistic regression).
Chapter 2. Related work

2.1.5 Co-evolutionary

Or adaptive, or STS Models ...

"Adaptive network is the term given to the types of network whose structure varies depending on the dynamics of the units on the nodes." according to Gross and Sayama [17]

Even if there are still properties to be discovered, the topology has been largely studied, as well as the dynamics on networks (sometimes by different fields, without the proper tools of network science), but the study of the effect of one on the other is recent.

The previous models were essentially built to study either the topology of a network, or how the nodes behave independently. But recently a new field of research emerged: the study of the networks with a mutual interaction between the state and the topology, sometimes called adaptive or co-evolutionary networks. For this work we will differentiate the two, we will call "adaptive" networks whose structures’ evolution exhibit a dependence toward their state (without necessarily a reciprocity), and "co-evolutionary"
2.1. Co-evolutionary classification

networks with a mutual influence of the state and the topology over each others’ evolution.

It has already been shown that topological properties have an effect on the dynamics, like in an epidemic event, but the opposite effect can also be considered: a road network will have new roads if it’s too congested (feedback loop between topology and nodes’ dynamics). The models able to exhibit a feedback loop between the state and the topology are called co-evolutionary networks models, as shown by Gross and Blasius [9].

As detailed in a previous section, the creation of a model which takes into account the state of the network to see the effect on the topology has been attempted by Kim and Leskovec [18], Dos Santos, De Almeida, Mendes, et al. [7] as well as Lee, Stephan, and Smola [8], but the first only study the state on the topology, whereas the second added the reflexive loop on the topological evolution, and the third propose a model which learn how to “interpret” the state to adapt the modification of the topology, without modifying the state in itself.

Co-evolution has rarely been studied, not from a computer scientist point of view: some physicists like SEUFERT and SCHWEITZER [23] contributed to such models, but is not described in terms of network generation, and does not provide social networks characteristics such as heavy-tail DD etc (because very specific to origin of life); therefore it’s not relevant for our work.

STS Classic

Kimura and Hayakawa [21] developed as well a model with attributes, which includes heterophily and homophily, but no preferential attachment.

STAS

---

He also called them adaptive networks, but in another paper he specifies that the adaptive networks don’t necessarily have the full loop from the topology to the state, so we will differentiate the adaptive and the co-evolutionary networks.
Chapter 2. Related work

SATS

From our field, Sayama, Pestov, Schmidt, et al. [22] proposed a Generative Network Automata with a real co-evolution between the state and the topology, but unfortunately it’s not relevant for this work, because it doesn’t provide a power-law DD (not representative of social networks).

SATAS

2.1.6 Details of highly relevant papers

As this work is mainly based on two papers, a deeper analysis is made:

DosSantos

[7] is the sequel of a previous paper, [24], which presented the scale-free homophilic network model.

This version add the notion of accentuation on the affinity term, as well as new analysis concerning the effect of the latter on a network.

In particular they observe that the clustering coefficient is stable for $n > 100$ (between 0.001 and 0.0001 depending on $m$, the number of edges added each iteration).

Another main observation is the scale-free character of the networks produced, with a power-law exponent around 2.9 (and correlated with $\sigma$ : the degree distribution is weakly influenced by the value of $\sigma$.)

They therefore introduce a new way to measure correlations between websites, continuing the work of [25].

A comparison with the Barabasi-Albert and the fitness model is made, and indeed this work is intended for websites, which partly explain why this model is not coevolutionary.

They invite to explore the impact of the variation of $\sigma$ ("For further work, the understanding of how $\sigma$ and different distributions of intrinsic characteristics supports a power-law $P(k)$ and its implications are a formidable challenge"), which we effectively tried to do, although in a different way.
Lee [8] was also a major influence for this present work.

The main hypothesis behind this paper is that even if preferential attachment helps understanding various network features, it fails when it comes to modeling the affinity between nodes ("preferential attachment models successfully capture key properties such as the degree distribution, they fail to include intuitive reasons behind the link generation: the affinity between vertices based on their latent or observed attributes"), thereby they strive to model that missing part of the link generation.

Their work gives similar results to real-life networks in many ways ("Experiments show that our model provides an excellent fit of many natural graph statistics and we provide an algorithm to infer the associated affinity function efficient"), such as the scale-free feature of the networks (they also prove that the power-law exponent is 2.48 "We will prove that the generated graphs show a power-law in degree distribution with exponent $2 + f(\Delta, t)$.").

The community structure has also been given attention to, and although they don’t provide any statistics, they show graphically that their model follows the same behavior as real-life graphs (as it happens, the Autonomous systems graph of Stanford).

As the previous paper, this model is intended for websites, and not for social networks. However we based our model predominantly on this one (the notion of sum in the node’s choice formula), as explained in a further section, but we also see and acknowledge the limitations from it: the lack of flexibility due to the nonexistence of a parameter balancing the affinity and the degree in the formula, as well as the normalization problem (here the degree is unbalanced compared to the affinity: it might be out of magnitude sometimes, when this subproblem is crucial in the study of homophily in social networks).
Chapter 3

The model

3.1 Construction

3.1.1 Why a model?

To understand phenomena, behaviors, and systems, we sometimes need a mathematical model to bring out the underlying mechanisms of these latter, or to check hypotheses. Indeed, even with good empirical observations, we might not be able to conclude something without trying to reproduce a simpler version of reality through experiments. Models, if correctly thought and implemented, can provide such an environment, and if a model with simple laws can fit a set of empirical data (on certain measures), it is not aberrant to think that those laws are the main mechanisms giving those measures. Models can’t really prove anything, but they can get us closer to a certain truth.

In network science, models are usually algorithms generating a set of nodes and edges, eventually attributed and weighted (respectively).

As mentioned earlier, existing network models were insufficient for what we wanted to investigate: they are either too focused on the topology, or on the state, whereas what we needed here is a network model which study the mutual influence of both. A model that exhibits both dynamics is called a co-evolutionary model: it has a feedback loop between the state and the topology.

Thinking in terms of co-evolutionary networks allows us to study the relation between the structure and what’s inside each node when the network is evolving.

But even the few recent models including both dynamics are not fit for SNA, and especially the study of opinion, so we had to create another one, more complete, and more suited to this project.

3.1.2 Tools

After the theoretical conception, we decided to implement it with an open-source Python software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks: https://networkx.github.io/.
3.2 Our implementation

To present the model, we will detail three main components:

- The core, which doesn’t have to change
- The choice of the new nodes, we investigated several ways to do so
- The way we update the state, probably the most important part, but we had less time to dig into that direction

3.2.1 Core

Graph $G(V,E,\eta)$

$V$ set of vertexes

$E$ set of edges

$\eta$ set of attributes

$m \in \mathbb{R} \setminus \{0\}$ number of edges added each step

$d(v)$ degree of node $v$

$A_{i,j}$ similarity between nodes $v_i$ and $v_j$, $f(v_i, v_j)$ similarity function between the two nodes where $0 < A_{i,j} \leq 1$ and $A_{i,j} = 1$ means that the two nodes are dissimilar, and $A_{i,j} = 1$ means that they’re very similar.

To a graph with $n \in \mathbb{N} > m$ nodes, each step we add a node $v_{n+1}$, to this node we connect $m$ edges. This connect to an existing vertex $v_e$ with a probability described in the next section.

3.2.2 Choice of node

We grouped the way to choose a node in two main categories (themselves composed of several sub-categories):

- The product formula
- The sum formula

$P_t$ is the abstract notion of topology, or how the degree impact the choice of a node (see below for the practical aspects).

$P_s$ is the notion of state, or how the affinity between the new node and the candidate node impact the choice of the latter.

**Product**

Each node will have the following weight to be chosen (which will be normalized according to all weights to give a probability between 0 and 1)

$$P\{(v_{n+1}, v_e)\} = P_s(v_{n+1}, v_e) \times P_t(v_{n+1}, v_e)$$

To avoid a null affinity when two nodes are too dissimilar, we had a parameter $\epsilon$ to avoid this case:

$$A_{i,j} = \frac{f(v_i, v_j) + \epsilon}{1 + \epsilon}$$

$\epsilon$ has to be very small
DosSantos2015 We tested the previous theoretical choice with the implementation of [7]

\[ P_s(v_{n+1}, v_e) = A_{n+1,e}^\sigma \]
\[ P_t(v_{n+1}, v_e) = d(v_e) \]

With this version, we can obtain a polarized network with a \( \sigma \) big enough (the definition of \textit{big enough} depends on the size of the network, how we update the state etc).

But as we increase the tuning parameter, we come from a power-law degree distribution to a geometric degree distribution (similar to model A of the BA model see [26]).

It is not possible to completely remove the degree (with a high sigma, between two nodes with a similar affinity, the new node will most likely choose a high degree node).

Conversely it can’t connect to high degree nodes if they are not similar enough.

\( \sigma = 0 \) means that we don’t take into account the affinity, but as it increases new nodes will tend to favor similar ones, but will most likely choose a highly connected node (among the range of similar nodes).

After several experimentations, we decided not to continue with this version, the main reason being that we can’t completely remove the impact of the topology (the degree) in the formula.

Sum

Here as well, each node will have the following weight to be chosen (which will be normalized according to all weights to give a probability between 0 and 1)

\[ P\{v_{n+1}, v_e\} = \theta P_t(v_{n+1}, v_e) + (\theta - 1)P_s(v_{n+1}, v_e) \]

where : \( \theta \in \mathbb{R} \) and \( 0 < \theta < 1 \)

We kept \( A_{i,j} = \frac{f(v_i, v_j) + \epsilon}{1 + \epsilon} \)

Lee2015 We tested the previous theoretical choice with the implementation of [8], here they don’t let the possibility to balance between the affinity and the degree.

\[ P\{v_{n+1}, v_e\} = d(v_e) + A_{n+1,e} - 1 \] And normalization

But as no balancing is possible, we decided to add a new parameter (\( \theta \)).
Chapter 3. The model

**Classic**

\[ P\{v_{n+1}, v_e\} = \theta P_t(v_{n+1}, v_e) + (\theta - 1) P_s(v_{n+1}, v_e) \]

where \( P_t(v_{n+1}, v_e) = d(v_e) \) and \( P_s(v_{n+1}, v_e) = A_{n+1,e} \)

Unlike the previous implementation, here we can balance the impact topology and the state: the differentiation is clear.

But the main technical problem is the normalization: the affinity is bounded between 0 and 1, whereas the degree is bounded between \( m \) (the number of edges we had each turn) and a large number, immediately dropping the impact of the affinity in the formula.

And even if we can reach the equilibrium between the two components (for a fixed size of graph), we observed that the system will not tend to connect similar nodes: it will tend to connect them a bit, but this is not significant (the average deviation of attributes, for the 1-hop and 2-hops neighbors, is the same as with a random graph).

We investigated a new \( P_t(v_{n+1}, v_e) = \frac{d(v_e)}{\sum_{i=1}^{n} d(v_i)} \), but on this case it's the degree which is minimized.

Here, unlike the previous version, the affinity is not normalized, to try to balance between the affinity and the degree.

It seemed better in the first place, but the degree was minimized by the affinity (we ended up losing the power-law degree distribution).

The main advantage is the tractability.

**Mixed** Based on the previous theoretical investigations, and the experimental results, we tried a mixed formula, with either the concept of differentiation through \( \theta \), but also the increase of affinity through \( \sigma \), as well as a new normalization for the topology:

\[ P_t(v_{n+1}, v_e) = \frac{d(v_e)}{\max(d(v_i))} \]

\[ P_s(v_{n+1}, v_e) = A_{n+1,e}^\sigma \]

\( \theta \) offers the tractability of choosing either the topology or the affinity, whereas \( \sigma \) can press the impact of the affinity to force the network to connect nodes only if they are very similar (unlike the classic sum formula).

### 3.2.3 Summary Table

See table 3.1
### 3.2. Our implementation

#### Product Sum Formula

\[
P_s(v_{n+1}, v_e) \times P_t(v_{n+1}, v_e) = \theta P_t(v_{n+1}, v_e) + (\theta - 1) P_s(v_{n+1}, v_e)
\]

<table>
<thead>
<tr>
<th>Tractability affinity</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tractability degree</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Pros</td>
<td>Easier to keep a balanced probability between a similar and a high degree node</td>
<td>Better differentiation</td>
</tr>
<tr>
<td>Cons</td>
<td>Can’t remove the degree, conversely can’t connect to high degree node if not similar enough</td>
<td>Without the right normalization, one of the two components can be negligible compared to the other</td>
</tr>
<tr>
<td>Polarization</td>
<td>Yes (easy to obtain)</td>
<td>Yes (less obvious)</td>
</tr>
<tr>
<td>Observations</td>
<td>The average deviation of attributes is controlled by ( \sigma )</td>
<td>High average deviation of attributes (the same with a random graph) without ( \sigma )</td>
</tr>
<tr>
<td>Varying Parameter</td>
<td>( \sigma = 0 ) no affinity, as it grows it tends to favor similar nodes, but will most likely choose a highly connected node among those</td>
<td>( \theta = 0 ) only topology ( \theta = 1 ) only affinity</td>
</tr>
<tr>
<td>Balance Probability</td>
<td>see A</td>
<td>see A</td>
</tr>
<tr>
<td>Other</td>
<td>( \sigma ) favors similar nodes</td>
<td>Edge effect ( \theta \to 1 )</td>
</tr>
<tr>
<td>Peculiar case 1</td>
<td>see A</td>
<td>see A</td>
</tr>
<tr>
<td>Peculiar case 2</td>
<td>see A</td>
<td>see A</td>
</tr>
<tr>
<td>Peculiar case 3</td>
<td>see A</td>
<td>see A</td>
</tr>
</tbody>
</table>

Table 3.1: Comparison product and sum formulas: summary table
3.2.4 State update

After the addition of each node, we update the state ($\eta$) of each node according to a breadth-first search algorithm.

For each $v_j$,

$$\eta_j' = \bar{\eta}_k$$

(\text{the mean of the neighborhood})

$k \in N_G(v)$, if the graph is oriented, then the neighborhood is reduced to the external one.

We observed that this update ruins any further observations, as all nodes will smooth their affinities according to their neighbors’, and this extremely fast. The value of the attribute $\eta$ will converge toward 0.5, giving Gaussian curve for the distribution of attributes.

We attempted to reduce this convergence, so that the nodes will slowly adapt to their neighbors’ opinions.

$$\eta_j' = \sqrt{|\eta_j - \bar{\eta}_k|}$$

It stems the convergence, but doesn’t stop it: this step is crucial but still to be detailed.

3.3 Explanation of the parameters

In our python implementation, we used different parameters, in the following section we will endeavor to explain them, as well as their effect.

3.3.1 Of the importance of initialization

As largely discussed in the scientific community [27], chaotic systems are extremely sensitive to initial conditions, so we tried different starts to study an eventual impact.

Different starts

We studied 5 different start configuration:

- A node alone, with random values
- A complete graph of $m$ nodes, with random values
- A complete graph of $2m$ nodes, $m$ with an intrinsic $\eta$ of 0, $m$ with 1
- A complete graph of $m$ nodes, with an intrinsic $\eta$ of 0
- A complete graph of $m$ nodes, with an intrinsic $\eta$ of 1

All of those different starts don’t influence the rest of the experiment, at least if we have a decent number of nodes (>50).
3.3. Explanation of the parameters

Choice of attributes

In addition to the attribution of fixed and random values to the initial clique, we analyzed two ways to initialize a new node:

- With a random value between 0 and 1
- With a random value of either 0 or 1

Once again, those two variants don’t impact the general behavior, most likely because the update of the state smooths the attribute distribution. We therefore advise a new investigation after the set up of a non-destructive update.

3.3.2 Topology update

Number of edges and nodes

The number of nodes doesn’t seem to impact anything, within reasonable bounds (>50 nodes). He haven’t tested any experiment with more than 1000 nodes, but so far we didn’t observe any change of behavior beyond 100 nodes. The same remark applies to the number of edges added each turn: the system behaves linearly against it.

Tuning $\sigma$

In the DosSantos as in the mixed formula, we let ourselves the possibility to tune the parameter $\sigma$, corresponding to the amount of importance we give to affinity in the choice of a node. It appears that over a certain value\(^1\), we observe a polarization of the graph (see chapter 4): the network is visually split into two very connected communities, resulting in a bimodal attribute distribution. We will discuss our measurements in a following section.

Tuning $\theta$

In the sum as in the mixed formula, we let ourselves the possibility to tune the parameter $\theta$, corresponding to the weight allocated to the degree or the affinity in the choice of a node.

Because of several unfixed bias (notably the state update), we couldn’t observe anything relevant.

3.3.3 State update

We believe that this step is crucial, but so far we couldn’t produce a significant update for our model. Indeed a quick update can negate the effects of most parameters, preventing the observation of a potentially relevant event.

\(^1\)With a state update quickly adapting, and 200 nodes, at a threshold of 12
Speed of convergence

We considered two speeds at which the update can take place:

- A quick one (corresponding to $\eta'_j = \bar{\eta}_k$)
- A slower one ($\eta'_j = \sqrt{|\eta_j - \bar{\eta}_k|}$)

So far we couldn’t notice anything significantly different, the second version only slowing down the rate at which the attribute distribution converge toward a mean value. We strongly think that in order to obtain a more realistic model, this step has to be dug, and so in a radically different way.

Weight of convergence

We considered two ways to consider the neighborhood when updating the state:

- Compute $\bar{\eta}_k$ without weight
- Compute $\bar{\eta}_k$ with weights (corresponding to the respective degree of each node)

Once more, no relevant observations were made.

3.4 Hypothesis

From the first theoretical analysis and a few experiments, we were able to extract several hypotheses:

**Hypothesis 1.** The emergence of polarization can only appear if we combine a strong homophily in the network, and a co-evolutionary behavior of the model.

**Hypothesis 2.** The state update is a major step, therefore the speed of adaptation must be subtly implemented in order to save the possibility to observe events such as creation of communities, differentiation of opinions etc.

**Hypothesis 3.** Pressing the affinity can lead to a progressive loss of the scale-free feature of a network, due to the diminution of the importance given to the $P_t$.

**Hypothesis 4.** By tuning the $\theta$ or the $\sigma$ parameter, a phase transition should appear by plotting one of the measure devised in chapter 4 against the tuning parameter.

These hypotheses will be challenged in the next chapter.
Chapter 4

Experiments

We ran several simulations to see if our main hypothesis was validated or invalidated, where we varied settings of the model. The results were put on perspective, and compared to a real dataset.

4.1 Datasets

4.1.1 Political blog

We used a dataset recorded in 2005 by Adamic and Glance of a directed network of hyper-links between web blogs on US politics. Adamic and Glance [28]

This network is composed of 1490 vertices and 19090 edges, the vertices represent the blogs, annotated with 0 (left or liberal) and 1 (right or conservative).

We used the graphml version of the dataset.

4.2 Analysis Tools

To analyze both the dataset and our generated networks, we used NetworkX. We worked with measures already implemented in the software, and also our own measures such as :

4.2.1 Average Attribute Deviation

The Average Attribute Deviation (AAD) is the standard deviation of an attribute for each node and their neighborhood, averaged for the network. This is our main measure for homophily : we observed values around 0.25 for a random network, and from 0.15 to 0.05 for homophilic networks. As we couldn’t find any references for this measure, we propose it as an original contribution.

4.2.2 Bimodality

The Bimodality Coefficient (BC), computed on the distribution of attributes, can give us an idea of how polarized the network is, because it should return a high number (close to 1) for a distribution highly bimodal, and a low (close to 0) if it’s not the case.
We computed the BC on the attribute distribution smoothed with the Gaussian Kernel Density Estimation (KDE) provided by the scipy package.

After several experiments, we realized that this measure was not reliable: indeed, a close to uniform distribution can have a high BC, whereas a visually bimodal distribution doesn’t necessarily have a high BC.

Thus we investigated the Bimodality Amplitude (BA) of Zhang, Mapes, and Soden [29], which is not suppose to return something superior to 0 without the presence of an effective bimodal distribution (also on the KDE).

Once again, we observed that in practice this measure is heavily biased. This bias is due to the fact that a uniform distribution, or at least something very close to it, can have a slight curve fooling the KDE, see figures 4.1 and 4.2.

Those two measure combined $BA \times BC$ give us the Bimodality Index
4.2. Analysis Tools

Figure 4.2: Attribute distribution of a network with state update, the blue curve is the KDE. We can see the obvious polarization of the network.

(BI), more robust that the BC, but still to be improved, eventually with a more satisfactory measure of the amplitude.

4.2.3 Power-law Coefficient

The powerlaw package of python provides functions to fit a distribution into a power-law, notably on discrete data, and can return the coefficient of the latter.

As the function will always return a coefficient (even if the distribution is not scale-free), we can distinguish a real scale-free distribution from a random one if the coefficient is probable (around 3 according to Barabasi), or not (some of our networks were fitted with a coefficient of 6 or more).
4.2.4 Polarization

Beside the bimodality, we also computed different measures of polarization. The first one was found on an online thread. This measure presents the same problems as presented above in the bimodality paragraph.

We also implemented the polarization index of [14], this measure presents the same imperfections, but has the advantage of taking into account the size of the two poles: a balance is required to have a truly polarized network. The BI seems to be the most reliable measure of polarization so far (over a certain threshold), but this area has to be defined better.

4.2.5 Others

Other measures were investigated, such as the Clustering Coefficient, the number of n-cliques, the number and the size of the communities, but they were not very meaningful for our purpose, nor reliable.

4.3 Results of the experiments

The results presented thereafter have been extracted from the mining of networks with $n = 100, m=2$, the mixed formula 3.2.2, and reproduced between 35 and 70 times for more precision. The results are sorted by the hypothesis they challenge.

4.3.1 Hypothesis 1

As explained in 2.1.6, the model of [8] is not very flexible, and the contrast in the order of magnitude between the preferential attachment and the affinity term is too big: they are not exhibiting a particular behavior, mostly correlating evidences.

Indeed, they are presenting a model which take into account affinity between agents, but the result cannot be an homophilic network.

We show that the simple fact of tuning the affinity parameter results in a network becoming more and more homophilic, see figure 4.3:
4.3. Results of the experiments

![Figure 4.3: Networks without state update, showing the gain in homophily](image_url)

But without state update, the model keeps being non-polarised, see figure 4.4:
Figure 4.4: Networks without state update, showing no polarization. This measure, although the best we currently have, is biased, as non-polarized networks shouldn’t exhibit something else than 0. See chapter 4.2

The emergence of homophily also happens when we add the state update part in the model, regardless the tuning of the affinity: the ADD is around 0.06, despite the strange behavior of the curve, the variation is very low (compared to the non-coevolutionary version), see figures 4.5a and 4.5b.
4.3. Results of the experiments

(A) Networks with state update, showing the strong homophily

(B) Networks with state update, showing the strong homophily, with a zoom on theta close to 1

**Figure 4.5:** Networks with state update, showing the strong homophily
This result can seem obvious (if the agents share a part of their information, they will most likely tend to be more similar), but an interesting thing appears when we combine both the possibility to press the affinity to make a network more homophilic, as well as the state update: the network starts to be polarized, see figures 4.7a and 4.7b.

![Attribute distribution of the dataset](image)

**Figure 4.6**: Attribute distribution of the dataset

We observe values close to those observed in a network without state update, the bias was treated in section 4.2.

Our results correlate the one observed in the dataset: we found an ADD of 0.11 (for a network we know that is polarized), and even if our measures doesn’t apply on non-continuous data (see figure 4.6), this dataset obviously possesses a high polarization: two poles, with a similar size, each containing a very different opinion (here either liberal or progressive).

$\sigma$ doesn’t seem to have an impact here, or at least not a very significant one.
4.3. Results of the experiments

(A) Networks with state update, showing the emergence of polarization

(B) Networks with state update, showing the emergence of polarization, with a zoom on theta close to 1

**Figure 4.7:** Networks with state update, showing the emergence of polarization
4.3.2 Hypothesis 2

As expected, a wrong balance of the state update can lead to the impossibility of observing events, all the network converging to an average opinion, see figure 4.8.

![Figure 4.8: Attribute distribution with a too fast adaptation](image-url)
4.3. Results of the experiments

4.3.3 Hypothesis 3

As aforementioned above, we also note that the networks generated tend to lose their scale-free feature as we press the affinity, see figure 4.9.

Figure 4.9: Networks with state update, showing the loss of the scale-free feature.

This is interesting because as showed in [30], "On non-polarized contexts, we observed a concentration of popular nodes along the boundary, since the sharing of similarities between members of the boundary increase the popularity of such nodes", whereas "polarized networks tend to have a lower concentration of popular nodes in the boundary, since the antagonism between both sides decrease the likelihood of existence of nodes that are popular in both groups".

An interpretation of the progressive loss of the scale-free character of
the networks would be that because of the creation of poles, we can’t observe highly-connected nodes.

The analysis of the degree distribution of the dataset gives us a power-law coefficient of 3.89, a high number for a social network, which might be symptomatic of a fragile scale free feature, or of different power-laws for each pole. This conclusion has to be confirmed.

We observed that the rest of the measures don’t vary, except the coefficient of power-law: we start to lose the scale free feature of the network, see figure 4.10

![Figure 4.10: Networks without state update, showing the loss of the scale-free feature](image-url)
4.3.4 Hypothesis 4

Contrary to what one might expect, we could not observe any real phase transition by varying a tuning parameter and then plotting various measures against this tuning parameter. So far we can consider this hypothesis as invalidated.

4.4 Analysis

Our hypothesis seems to be confirmed: we need the possibility to emphasize the affinity between people and the fact that they exchange a feedback to observe the creation of polarization.

The common sense explanation is: if you are with similar people on a social network, you’ll tend to be in a bubble of people sharing the same ideology ("Recent research has shown that the most prominent and politically active users mainly interact with their own partisans [29–31], leaving little space for real debate and cross ideological interactions." [14]), which leads to the creation of poles, extreme in their way of thinking and hermetic with respect to the other.

Bourdieu, probably one of the most important modern sociologist, defined structuralism in Bourdieu [31] as following:

*By structuralism or structuralist, I mean that there exist, within the social world itself and not only within symbolic systems (language, myths, etc.), objective structures independent of the consciousness and will of agents, which are capable of guiding and constraining their practices or their representations.*

Maybe the realization of the structures and of how they influence the agents is a first step towards the cessation of extremisation.
Chapter 5

Discussion and further work

5.1 Conclusion

We saw that polarization of opinions appears only if we have homophily of attributes and coevolution. But as social networks have been proved to be homophilic (MIPHerson, Smith-lovin, and Cook [32]), and coevolutionary (to an extent not measured yet), we can deduce that polarization emerges only with very homophilic networks, and that this phenomenon leads to a progressive loss of the scale-free character.

5.2 Limitations

5.2.1 Conceptual limitations

For this thesis, we chose to represent opinions as a number between 0 and 1, but a point of view on a topic cannot always be expressed as a single number, it cannot always be reduced to a binary choice, or even a distribution only between two dominant opinions. A more accurate numeric representation could be as a vector of opinion, each dimension of that vector corresponding to a fact of the opinion on a topic.

5.2.2 Technical limitations

The algorithm developed for the implementation of the model possesses a high complexity, therefore in its current state the model is not very scalable yet. Improving this drawback would need a restructuring of the data structure (the affinity matrix, and its numerous updates, contributes greatly to the cost).

5.3 Possible enhancements

In addition to the first four hypotheses, we can formulate one more, that could be tested in future works:

5.3.1 Hypothesis 5

Hypothesis 5. All the individuals are not influenced at the same pace, and to the same extent.

After testing, if this hypothesis turns out to be true, an idea of implementation could be to use DeGroot learning, see DeGroot [33]
We can also identify several weaknesses that would need to be improved, such as:

### 5.3.2 Measure of polarization

The study and implementation of a good and reliable measure of polarization, in order to efficiently determine how polarized is a network of opinion.

### 5.3.3 Working with text

The initial goal of this thesis was to work with text-networks: where each node is attributed with a piece of text (or its representation). The impact of the text on the evolution of a network has so far not been evaluated, and conversely how much does the topology of a network influences the text exchanged on the latter. For that a review of two categories of tools has been done: topic models and natural language processing tools.

#### Topic Models

In order to measure the similarity between two agents in a network, a useful way is to use Topic Models: they are statistical models displaying the topic behind a corpus a document. They compute the occurrences of words in a corpus, and assign a set of topics to each document, according to the words it is composed of. One can therefore compute the similarity between documents, according to the topics they belong to.

**Dirichlet Allocation** The LDA model of Blei, Ng, and Jordan [34] is probably the most used topic model, for two main reasons: it is quite simple, thus tractable and extensible, and also because it was one of the first (early 2000) and it gives good results despite its simple implementation.

**Pachinko Allocation Model** The PAM of Li and McCallum [35] works according to the same principle as LDA, but add a layer of abstraction: it correlates not only the words across a corpus of documents, but also the topics created beforehand by the first part of the algorithm. Where LDA captures correlations between pairs of topics, the PAM can capture correlations between $n$ topics.

**Author-Topic Model** The ATP of Rosen-Zvi, Griffiths, Steyvers, et al. [36] is an extension of LDA to include authorship: each author is associated with a multinomial distribution over topics and each topic is associated with a multinomial distribution over words.

**Twitter-Network Topic Model** The TN of Lim, Chen, and Buntine [37] enables additional informative inference such as authors’ interests, hashtag analysis etc. It is based on a nonparametric Bayesian approach, and was conceived more specifically for twitter.
5.3. Possible enhancements

**Distance Dependent Chinese Restaurant Processes**  The DPCRP of Blei and Frazier [20] is used by Lee, Stephan, and Smola [8]

**Natural Language Processing**

To produce efficient Topic Models, we need Natural Language Processing (NLP) to parse the documents required to create our topics, and also if only to compute efficient measures on the text (not necessarily with topics).

Indeed, in order to remove typos, stop-words (such as *the, is, at* etc, usually not relevant), to determine the language used in a message,

**TwitIE**  TwitIE by Bontcheva, Derczynski, Funk, et al. [38] is a NLP pipeline customised for microblog text, intended to properly extract information from tweets.

**LDbin**  The LDbin of Lui and Baldwin [39] is a method for language identification, fast and reliable.

**Linguistic Homophily**  Yang and Eisenstein [40] developed a method using language homophily (the tendency of socially linked individuals to use language similarly) to create sentiment analysis models.

**Bad language**  An efficient part-of-speech tagging (POST, marking up a word in a text as corresponding to a particular part of speech such as nouns, verbs, adjectives etc) is not perforce one which will remove all *dirty* data (grammatically wrong), as proved by Eisenstein [41].
Appendix A

Comparison: product vs sum

We will consider the product formula as \([7]\), and the classic sum formula. This bias is crucial in the analysis of the next appendix.

A.1 Peculiar case 1

A.1.1 Description

We have \(n\) nodes with \(\eta = 0\), 1 node with \(\eta = 1\), all the 0 nodes are connected to the 1 node.
\(d(v_0) = 1\) and \(d(v_1) = n\).
A new node with \(\eta = 0\) wants to connect, it can either choose the highly connected but extremely dissimilar node, or one of the identical but lowly connected node.

A.1.2 Product behavior

\(P(0) = 1\) and \(P(1) = 0\)
After normalization:
\(n\) nodes with \(p = 1/n\), 0.01 for \(n = 100\).
1 node with \(p = 0\)
   The highly connected node will never be chosen.

A.1.3 Sum behavior

\(P(0) = 1 + 1 = 2\) and \(P(1) = 0 + n = n\)
After normalization:
\(n\) nodes with \(p = 2/(2n + n) = 2/3n\) 0.007 for \(n = 100\).
1 node with \(p = n/3n = 1/3\) 0.33 for \(n = 100\).

Here, the highly connected node has a clear advantage.

A.2 Peculiar case 2

A.2.1 Description

We have \(n\) nodes with \(\eta = 0\), 1 node with \(\eta = 1\), all the 0 nodes are connected to the 1 node.
\(d(v_0) = 1\) and \(d(v_1) = n\).
A new node with \(\eta = 0.1\) wants to connect, it can either choose the highly
connected but dissimilar node, or one of the similar but lowly connected node.

A.2.2 Product behavior

Low $\sigma$

$\sigma = 1$

$P(0) = 0.9$ and $P(1) = n/10$

After normalization:

n nodes with a $p = 0.9/n$ 0.01 for $n = 100$.

1 node with a $p = 10/n$ 0.01 for $n = 100$.

The highly connected node has a similar probability to be chosen as the others nodes.

High $\sigma$

$\sigma = 10$

$P(0) = 0.35$ and $P(1) = n \times 10^{-10}$

After normalization:

n nodes with a $p = 0.35/(0.35 \times n + n \times 10^{-10})$ 0.01 for $n = 100$.

1 node with a $p = n \times 10^{-10}/(0.35 \times n + n \times 10^{-10})$ 0 for $n = 100$.

The highly connected node will never be chosen.

A.2.3 Sum behavior

$P(0) = 0.9 + 1 = 1.9$ and $P(1) = 0.1 + n = n.1$

After normalization:

n nodes with a $p = 1.9/(1.9n + n.1) = 1.9/3n$, 0.007 for $n = 100$.

1 node with a $p = n.1/3n \approx 1/3 0.34$ for $n = 100$.

Here, the highly connected node has a clear advantage.

A.3 Peculiar case 3

Bunch of nodes with $\eta = 0$

Bunch of nodes with $\eta = 1$

Same distribution of degrees in each bunch. A new node with $\eta = 0.1$ wants to connect, it can either choose to similar nodes, or dissimilar nodes, in each case among one of the bunch.

A.3.1 Product behavior

Will only connect to very nodes in the similar bunch, among those nodes, it will probably choose a highly connected node (as the influence of the degree cannot be removed).
A.4. Equilibrium between probabilities

A.3.2 Sum behavior
Will connect to a highly connected node, regardless the affinity.

A.4 Equilibrium between probabilities

The goal here is to analyze the behavior of the two main formulas, considering a choice between either a similar but lowly connect node, or a dissimilar but highly connected node, to see how the balance between $P_s$ and $P_t$ can be reached, and under which circumstances.

Consider node $v_1$ with a degree $d_1$ (low) and node $v_2$ with a degree $d_2$ (high). A new node $v_e$ wants to connect, it has an affinity $A_1$ (high) with $v_1$ and an affinity $A_2$ (low) with $v_2$.

To have the same probability to connect to each node (which in real life would be: I’ll either connect to someone because that person is similar to me, or I’ll either connect to someone because that person is very popular), you need:

A.4.1 Product

$$(A_1/A_2)\sigma = d_2/d_1$$

A practical case where $A_1/A_2 = 9$ and $\sigma = 10$ would mean that even with a $d_1 = 1$, $d_2$ would need to be close to 3.5 billion, which is impossible to reach.

The point here is that if two nodes are too dissimilar, they will never connect to each other, even if the candidate node is extremely popular.

Conversely, if $A_1/A_2 \approx 1$, increasing $\sigma$ will not affect the choice (and the new node will most likely choose a highly connected node).

A.4.2 Sum

$$\theta(A_1 - A_2) = (\theta - 1)(d_2 - d_1)$$

The left part will tend to $\theta$ (if $A_1 \gg A_2$, then the difference will tend to 1), whereas the right part will tend to $(\theta - 1)d_2$ (if $d_2 \gg d_1$, then the difference will tend to $d_2$). We can therefore reduce the equation to:

$$d_2 \approx 1/(\theta - 1)$$

As $d_2 \to \infty$ we will have $\theta \to 1$

Where comes from the explanation of a difficult balance between the two components.
Bibliography


