Epidemic Processes on Networks, Viral Marketing and Optimal Vaccination

Alberto Guggiola
a.guggiola@gmail.com

École Normale Supérieure de Paris

NETADIS Scientific Kick-off Meeting
Torino (Italy), February 5th, 2013

Outline

Outline

Who I Am

- Bachelor Degree in *Physics* at the University of Torino
- Master Degree in *Physics of Complex Systems* at the University of Torino

Who I Am

- Bachelor Degree in *Physics* at the University of Torino
- Master Degree in *Physics of Complex Systems* at the University of Torino

- Now PhD student at the École Normale Supérieure de Paris

Outline

1. Presentation
2. Previous Research Projects
3. The NETADIS Project
4. Possible Secondments

Previous Research Projects

Bachelor Thesis

Spreading of Information on Dynamical Contact Networks

With **Ciro Cattuto**, ISI Foundation, Torino

Master Thesis

Statistical Sequence Analysis of Protein Families

With **Andrea Pagnani**, HuGeF Foundation, Torino

Previous Research Projects

Bachelor Thesis

Spreading of Information on Dynamical Contact Networks

With **Ciro Cattuto**, ISI Foundation, Torino

Master Thesis

Statistical Sequence Analysis of Protein Families

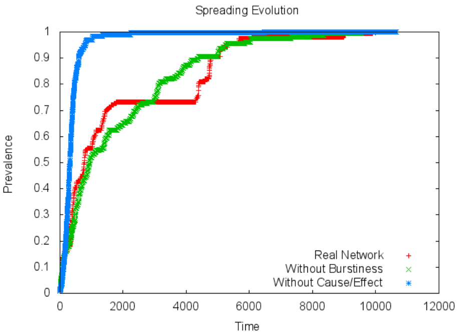
With **Andrea Pagnani**, HuGeF Foundation, Torino

Spreading of Information on Dynamical Contact Networks

- Real World networks:
  - Small world effect
  - Slow propagation of information

    M. Karsai et al., 2011: *Small But Slow World: How Network Topology and Burstiness Slow Down Spreading*, arXiv:1006.2125

- Which topological characteristics slow down the spreading?



Previous Research Projects

Bachelor Thesis

Spreading of Information on Dynamical Contact Networks

With **Ciro Cattuto**, ISI Foundation, Torino

Master Thesis

Statistical Sequence Analysis of Protein Families

With **Andrea Pagnani**, HuGeF Foundation, Torino

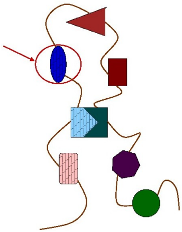
Statistical Sequence Analysis of PF

Objective

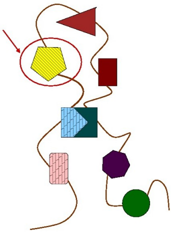
Inference of protein structures using statistics of the sequences

F. Morcos, A. Pagnani et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families, PNAS 2011

- Contact points are more correlated than generic ones



Original Sequence



Independent Modifications

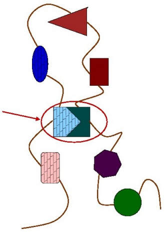
Statistical Sequence Analysis of PF

Objective

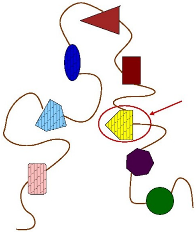
Inference of protein structures using statistics of the sequences

F. Morcos, A. Pagnani et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families, PNAS 2011

- Contact points are more correlated than generic ones



Original Sequence



One-site Modified Sequence

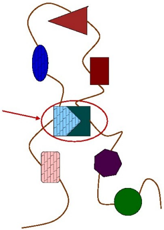
Statistical Sequence Analysis of PF

Objective

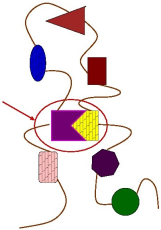
Inference of protein structures using statistics of the sequences

F. Morcos, A. Pagnani et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families, PNAS 2011

- Contact points are more correlated than generic ones



Original Sequence



Two-site Modified Sequence

Statistical Sequence Analysis of PF

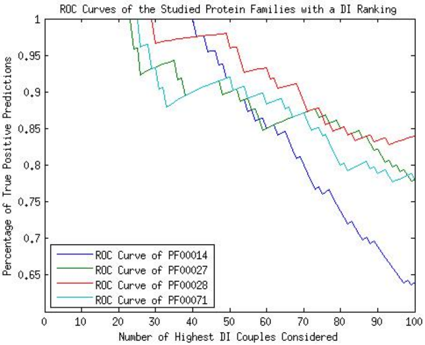
- *Direct Information* ranking of the couples of sites

Result

Percentage of TP among the x top ranked couples

Possible Improvement

Generate modified datasets, then rank according to mean DI



Statistical Sequence Analysis of PF

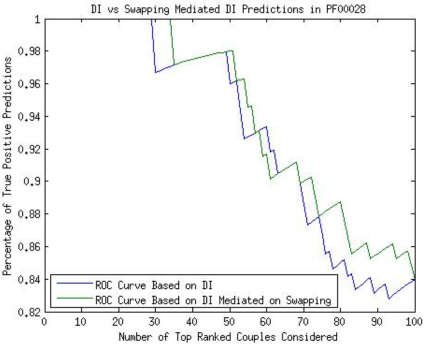
- *Direct Information* ranking of the couples of sites

Result

Percentage of TP among the x top ranked couples

Possible Improvement

Generate modified datasets, then rank according to mean DI



Outline

The NETADIS Project

Project 1 ⇝ Socio-economical sciences

With **Guilhem Semerjian**:

Epidemic processes on networks, viral marketing and optimal vaccination

Which initial conditions lead a system to a certain final state within a given dynamical process?

Project 2 ⇝ Biology

With **Rémi Monasson**:

Statistical mechanical study of the dynamics of interacting genomical sequences

PDZ proteic domains and their ligands are highly specific.
How is such an **optimal** result obtained through the two independent, local evolutions of the sequences?

The NETADIS Project

Project 1 ⇝ Socio-economical sciences

With **Guilhem Semerjian**:

Epidemic processes on networks, viral marketing and optimal vaccination

Which initial conditions lead a system to a certain final state within a given dynamical process?

Project 2 ⇝ Biology

With **Rémi Monasson**:

Statistical mechanical study of the dynamics of interacting genomical sequences

PDZ proteic domains and their ligands are highly specific.
How is such an **optimal** result obtained through the two independent, local evolutions of the sequences?

General Framework

Microscopically Irreversible Process

- **Binary variables** $x_i^t \in \{0, 1\}$ $i \in \{1, ..., N\}$ and $t \in \{0, ..., \infty\}$
- **State of the system at time t**: $\vec{x}^t = \{x_1^t, x_2^t, ..., x_N^t\}$
- **Only flips $0 \rightarrow 1$ allowed**: $x_i^t = 1 \rightarrow x_i^{t'} = 1 \quad \forall t' \geq t, \forall i$

Dynamics: Bootstrap Percolation

$$x_i^{t+1} = \begin{cases} 1 \text{ if } x_i^t = 1 \text{ or } \sum_{j \in \partial_i} x_j^t \geq \theta \\ 0 \text{ otherwise} \end{cases}$$

Applications

Financial contagion, failures in power grids, viral marketing ...

General Framework

Microscopically Irreversible Process

- **Binary variables** $x_i^t \in \{0,1\}$ $i \in \{1,...,N\}$ and $t \in \{0,...,\infty\}$
- **State of the system at time** t: $\vec{x}^t = \{x_1^t, x_2^t, ..., x_N^t\}$
- **Only flips** $0 \to 1$ **allowed**: $x_i^t = 1 \to x_i^{t'} = 1 \quad \forall t' \geq t, \forall i$

Dynamics: Bootstrap Percolation

$$x_i^{t+1} = \begin{cases} 1 \text{ if } x_i^t = 1 \text{ or } \sum_{j \in \partial_i} x_j^t \geq \theta \\ 0 \text{ otherwise} \end{cases}$$

Applications

Financial contagion, failures in power grids, viral marketing ...

General Framework

Microscopically Irreversible Process

- **Binary variables** $x_i^t \in \{0,1\}$ $i \in \{1, ..., N\}$ and $t \in \{0, ..., \infty\}$
- **State of the system at time** t: $\vec{x}^t = \{x_1^t, x_2^t, ..., x_N^t\}$
- **Only flips** $0 \to 1$ **allowed**: $x_i^t = 1 \to x_i^{t'} = 1 \quad \forall t' \geq t, \forall i$

Dynamics: Bootstrap Percolation

$$x_i^{t+1} = \begin{cases} 1 \text{ if } x_i^t = 1 \text{ or } \sum_{j \in \partial_i} x_j^t \geq \theta \\ 0 \text{ otherwise} \end{cases}$$

Applications

Financial contagion, failures in power grids, viral marketing ...

The Spread Optimization Problem

Statement

Given a graph $\mathcal{G} = \mathcal{G}(\mathcal{V}, \mathcal{E})$, find the minimal set of seeds such that, at a certain time T, all the nodes of the network become active

F. Altarelli, A. Braunstein, L. Dall'Asta, R. Zecchina, The Spread Optimization Problem, arXiv:1203.1426

Energy function

$$\varepsilon(\vec{t}) = \sum_{i=1}^{N} \varepsilon_i = \sum_{i=1}^{N} \left(\mu \mathbb{I}[t_i = 0] + \varepsilon \mathbb{I}[t_i = \infty] \right)$$

where t_i is the **activation time** of the i^{th} node s.t. $x_i^{t_i - 1} = 0$ and $x_i^{t_i} = 1$

Large Deviations

The seeds are **not** randomly chosen $\rightsquigarrow$ Extremal properties of the process under exam

The Spread Optimization Problem

Statement

Given a graph $\mathcal{G} = \mathcal{G}(\mathcal{V}, \mathcal{E})$, find the minimal set of seeds such that, at a certain time T, all the nodes of the network become active

F. Altarelli, A. Braunstein, L. Dall'Asta, R. Zecchina, The Spread Optimization Problem, arXiv:1203.1426

Energy function

$$\varepsilon(\vec{t}) = \sum_{i=1}^{N} \varepsilon_i = \sum_{i=1}^{N} \Big(\mu \mathbb{I}[t_i = 0] + \varepsilon \mathbb{I}[t_i = \infty] \Big)$$

where t_i is the **activation time** of the i^{th} node s.t. $x_i^{t_i - 1} = 0$ and $x_i^{t_i} = 1$

Large Deviations

The seeds are **not** randomly chosen $\rightsquigarrow$ Extremal properties of the process under exam

The Spread Optimization Problem

Statement

Given a graph $\mathcal{G} = \mathcal{G}(\mathcal{V}, \mathcal{E})$, find the minimal set of seeds such that, at a certain time T, all the nodes of the network become active

F. Altarelli, A. Braunstein, L. Dall'Asta, R. Zecchina, The Spread Optimization Problem, arXiv:1203.1426

Energy function

$$\varepsilon(\vec{t}) = \sum_{i=1}^{N} \varepsilon_i = \sum_{i=1}^{N} \Big(\mu \mathbb{I}[t_i = 0] + \varepsilon \mathbb{I}[t_i = \infty] \Big)$$

where t_i is the **activation time** of the i^{th} node s.t. $x_i^{t_i - 1} = 0$ and $x_i^{t_i} = 1$

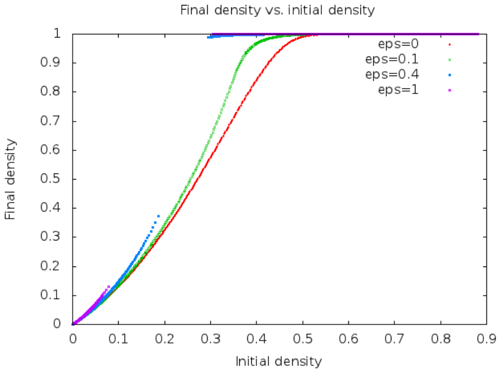
Large Deviations

The seeds are **not** randomly chosen $\rightsquigarrow$ Extremal properties of the process under exam

Optimization vs. Random Choice of the Seeds

- If $\varepsilon > 0$, the fraction of seeds needed to activate all the network is smaller with respect to the case where no optimization is imposed (i.e. $\varepsilon = 0$)

Random Regular Graph, K=3, $\Theta = 2$, $T = 20$



Phase Transitions in rCSP

Constraint Satisfaction Problems

A CSP consists in a set of constraints to be simultaneously satisfied by a set of variables.

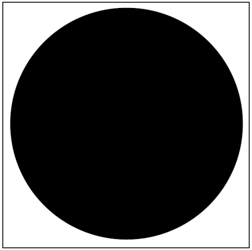
- As $\alpha = \frac{\#constraints}{\#variables}$ changes $\Rightarrow$ **Phase transitions**

PHASE TRANSITIONS IN rCSP

CONSTRAINT SATISFACTION PROBLEMS

A CSP consists in a set of constraints to be simultaneously satisfied by a set of variables.

- As $\alpha = \frac{\#constraints}{\#variables}$ changes $\Rightarrow$ **Phase transitions**



$\alpha = \alpha_S \Rightarrow$

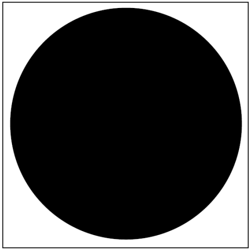
SAT-UNSAT TRANSI-TION

PHASE TRANSITIONS IN rCSP

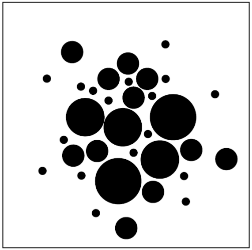
CONSTRAINT SATISFACTION PROBLEMS

A CSP consists in a set of constraints to be simultaneously satisfied by a set of variables.

- As $\alpha = \frac{\#constraints}{\#variables}$ changes $\Rightarrow$ **Phase transitions**



$\alpha = \alpha_C < \alpha_S \Rightarrow$ **CLUSTERING TRANSITION**



Beyond the RS Phase in the Spreading Problem

- For certain values of the parameters μ and ϵ the BP equations do not converge any more
- The hypothesis on which they hold are then no more fulfilled because of the **replica symmetry breaking**: the space of the solutions splits into separate pure states
- The problem should then be approached in a more general framework (i.e. **1 RSB**), so to study the statistics of sets of solutions instead of trying to find a unique one

$$\eta_{ij}(t_i, t_j) \quad \rightsquigarrow \quad P[\eta_{ij}(t_i, t_j)]$$

First Results and Perspectives

First Result

In the region where BP equations do not converge, the **complexity** (i.e. the logarithm of the number of pure states) has been seen to be strictly positive: **the 1 RSB approach seems then to be justified**

BUT a deeper analysis is needed in order to extract useful information about the spreading process

Outline

Possible Intersections with Other Projects

KCL1: Sub-network analysis using projection methods

What can be said about extreme trajectories in networks if the structure is partially unknown?

KCL2: Contagion dynamics across credit networks
ICTP: Inference in finance and socio-economic networks

Applications of the theory of spreading events developed to socio-economic data (epidemics, financial contagion...)

Thank you for your attention!!

Possible Intersections with Other Projects

KCL1: Sub-network analysis using projection methods

What can be said about extreme trajectories in networks if the structure is partially unknown?

KCL2: Contagion dynamics across credit networks
ICTP: Inference in finance and socio-economic networks

Applications of the theory of spreading events developed to socio-economic data (epidemics, financial contagion...)

Thank you for your attention!!