Minimal contagious sets in random regular graphs

Alberto Guggiola, Guilhem Semerjian

École Normale Supérieure de Paris

July 17^{th} 2014

Table of Contents

Definitions and applications

Epidemic process

Dynamical evolution of the states of the nodes in a graph, with contagion rules which depend on the state of their neighbours

Fields of applications: illnesses, economical systems, viral marketing

A recent example

(Controversial) paper on PNAS about *emotional contagion* on Facebook [1]

[1] Adam Kramer et al (2014), Experimental evidence of massive-scale emotional contagion through social networks, doi: 10.1073/pnas.1320040111

Different dynamics, different phenomena

Possible processes

- SI, SIS, SIR
- In particular, **monotonous** (SI, SIR) or **non-monotonous** (SIS)

Computationally hard problems to be studied

- Inference (e.g. patient zero in an infection)
- Optimisation [2]
  - minimal subset of nodes to vaccinate to stop a contagion
  - for fixed number of seeds, maximize the spreading
  - **minimum seed configuration activating all the network**

[2] F. Altarelli, A. Braunstein, L. Dall'Asta, and R. Zecchina, Journal of Statistical Mechanics: Theory and Experiment 2013, P09011 (2013)

The bootstrap percolation

At any given time t, a node i can be *active* ($\sigma_i^t = 1$) or *inactive* ($\sigma_i^t = 0$)

The evolution of the system

$$\sigma_i^{(t)} = \begin{cases} 1, & \text{if } \sigma_i^{(t-1)} = 1 \\ 1, & \text{if } \sigma_i^{(t-1)} = 0 \text{ and } \sum_{j \in \partial i} \sigma_j^{(t-1)} \geq l_i \\ 0, & \text{otherwise} \end{cases}$$

Main features

The process (a.k.a. **threshold model**) is **deterministic** and **monotonous**: $\underline{\sigma}^t = f(\underline{\sigma}^0)$

To have analytical results: $(k+1)$ RRG, $l_i \equiv l \; \forall i$

Random choice of the seeds on RRG

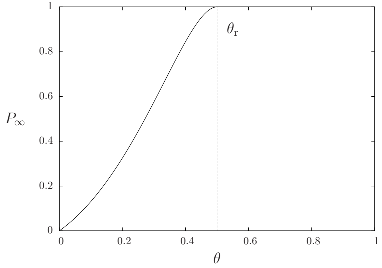
Def: threshold for random initial conditions

$$\theta_r(l,k) \text{ s.t. } \mathrm{Prob}(\sigma_{i_0}^{\infty}=1)|\theta \begin{cases} =1 & \text{if } \theta > \theta_r(k,l) \\ <1 & \text{if } \theta < \theta_r(k,l) \end{cases}$$

$k = l$

Continuous transition

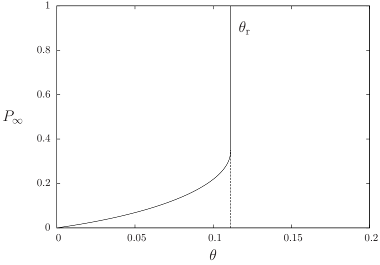
$\theta_r = \frac{k-1}{k}$



$k > l$

Discontinuous transition

No simple expression for θ_r



The optimisation problem

Def: Minimal density

$$\theta_{min}(G, \{l_i\}, T) = \frac{1}{N} \min_{\underline{\sigma}} \left\{ \sum_i \sigma_i | \sigma_i^T = 1 \forall i \right\}$$

- Original spreading maximization problem: $\theta_{min}(G, \{l_i\}, T = \infty)$
- Large deviation phenomenon: $\theta_{min} < \theta_r$

Mapping to other problems in graph theory

Arbitrary graph, $l_i = d_i \; \forall i$

Inactive sites correspond to a **maximum independent set** on the graph. Complete activation in one step.

$T = 1, \quad \forall l_i, d_i$

Biroli Mezard model: a site can be inactive if at most $d_i - l_i$ of its neighbors are inactive.

Mapping to other problems in graph theory

$T = \infty, \quad l_i = d_i - 1 \; \forall i$

Subgraph of inactive sites must be acyclic.
Seeds have to form a **decycling set** of the graph.

In $(k+1)$ regular graphs, $\theta_{min}(k,k) \geq \frac{k-1}{2k}$.
Known result[3]: $\theta_{min}(2,2) = \frac{1}{4}$; conjectured that $\theta_{min}(3,3) = \frac{1}{3}$

$T = \infty, \quad l_i < d_i - 1 \; \forall i$

Subgraph of inactive sites must not contain $d_i - l_i$ cores.
Seeds have to form a **"de-coring" set** of the graph.
Known result in regular graphs[4]: $\theta_{min}(k,l) \geq \frac{2l-k-1}{2l}$

[3] S.Bau,N.C.Wormald,S.Zhou,Random Structures & Algorithms **21**, 397 (2002)

[4] P.A.Dreyer,F.S.Roberts,Discrete Applied Mathematics **157**, 1615 (2009)

Statistical physics description

Probability measure over initial configurations

$$\eta(\underline{\sigma}) = \frac{1}{Z} e^{\sum_i [\mu\sigma_i - \varepsilon(1-\sigma_i^T)]} \xrightarrow{\varepsilon\to\infty} \eta(\underline{\sigma}) = \frac{1}{Z} e^{\mu \sum_i \sigma_i} \prod_i \mathbb{I}(\sigma_i^T = 1)$$

Minimal density

$$\theta_{min}(G, \{l_i\}, T) = \lim_{\mu\to-\infty} -\frac{1}{\mu}\frac{1}{N} \log Z(G, \{l_i\}, T, \mu, \varepsilon = +\infty)$$

More information

Def: entropy density $s(\theta)$

Number of percolating configurations with $\frac{1}{N}\sum_i \sigma_i^0 = \theta \sim e^{Ns(\theta)}$

Free-entropy density

$$\phi(G, \{l_i\}, T, \mu, \varepsilon = +\infty) = \sup_\theta[\mu\theta + s(\theta)] \Rightarrow \begin{cases} s(\theta) = \phi(\mu) - \mu\theta \\ \theta = \phi'(\mu) \end{cases}$$

Analogous relations also if $\varepsilon < \infty$

Table of Contents

The factor graph representation

The activation time description

$$t_i(\underline{\sigma}) = \min\{t \text{ s.t. } \sigma_i^t = 1\} \Rightarrow t_i(\underline{\sigma}) = f(\sigma_i, \{t_j(\underline{\sigma})\}_{j\in\partial i}; l_i)$$

- Local interactions
- Explicit σ_i^T

Duplication of the times

On each edge $\langle i, j\rangle$ a couple (t_{ij}, t_{ji}) of redundant variables is introduced

$$\eta(\underline{\sigma}, \underline{t}) = \frac{1}{Z}\prod_i w_i(\sigma_i, \{t_{ij}, t_{ji}\}_{j\in\partial i})$$

$$w_i(\sigma_i, \{t_{ij}, t_{ji}\}_{j\in\partial i}) = e^{\mu_i\sigma_i} e^{-\varepsilon_i \mathbb{I}(f(\sigma_i, \{t_{ki}\}_{k\in\partial i}; l_i) = +\infty)} \prod_{j\in\partial i} (t_{ij} = f(\sigma_i, \{t_{ki}\}_{k\in\partial i}; l_i))$$

A portion of the factor graph

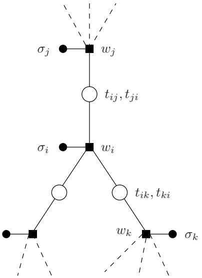


Figure: Factor graph of $\eta(\underline{\sigma}, \underline{t})$

Factor node $w_i \rightarrow$ interaction among σ_i and (t_{ij}, t_{ji}) for any $j \in \partial i$.

Table of Contents

Recursive computation on Z

RS ansatz is legitimate if the graph is a tree, or is tree-like with an assumption of long-range correlation decay.

Definition of the messages

For each directed edge $i \to j$ a **message** $\eta_{i\to j}(t_{ij}, t_{ji})$ (probability distribution over pair of activation times) is introduced

RS recursion

$$\eta_{i\to j} = g(\{\eta_{k\to i}\}_{k\in\partial i\setminus j}; l_i, \varepsilon_i, \mu_i)$$

From the converged values of the messages $\to$ thermodynamical quantities (e.g. ϕ)

A convenient reparametrisation

Each $\eta(t, t')$ is described by $(T+2)^2 - 1$ independent real numbers.

$$\Downarrow$$

Encodable in $h = (a_0, ..., a_T, b_{T-1}, ..., b_1)$ ($2T$ numbers)

Recursions among cavity fields

$$h = g(h_1, ..., h_k)$$

Single link approach

In RRG, **factorized solution**: $h_i \equiv h \; \forall i \Rightarrow h$ fixed point of $h = g(h, ..., h)$

Violation of RS ansatz

Apparently reasonable results...

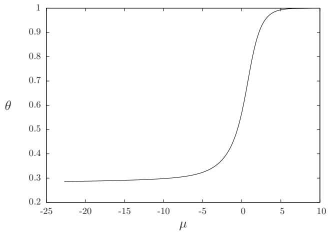


Figure: RS prediction of $\theta(\mu)$

... but unphysical predictions!

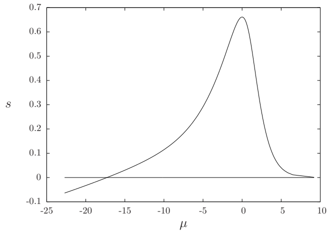


Figure: RS prediction of $s(\mu)$

RS Breaking

Constraints harder to satisfy $\Rightarrow$ **Long range correlations** among variables

Table of Contents

The 1RSB ansatz

Fragmentation of the configuration space into **clusters**, s.t. correlation decay inside each cluster γ.

Complexity function $\Sigma(\phi)$

Number of clusters with internal free-entropy density $\phi_\gamma \sim \phi \simeq e^{N\Sigma(\phi)}$

1RSB thermodynamical quantities

$$\Phi(m) = \frac{1}{N} \log \sum_\gamma Z_\gamma^m = \sup_\phi [\Sigma(\phi) + m\,\phi]$$

Parametrical reconstruction of $\Sigma(\phi)$

$$\Sigma(\phi_{int}(m)) = \Phi(m) - m\phi_{int}(m); \quad \phi_{int}(m) = \Phi'(m)$$

The 1RSB recursions

Single sample equations

On each directed edge: probability distribution $P_{i\to j} = G\left[\left\{P_{k\to i}\right\}_{k\in\partial i\setminus j}\right]$

$$P(h) =$$
$$= \frac{1}{Z_{iter}(P_1, ..., P_k)} \int dP_1(h_1)...dP_k(h_k)\delta(h-g(h_1, ..., h_k))z_{iter}(h_1, ..., h_k)^m$$

Single link approach

On RRG, factorized solution:

$$P(h) = \frac{1}{Z_{iter}} \int dP(h_1)...dP(h_k)\delta(h - g(h_1, ..., h_k))z_{iter}(h_1, ..., h_k)^m$$

Usual pattern in constraint satisfaction problems

RS phase

$\mu > \mu_d \Rightarrow$ No non-trivial solution of 1RSB equations for $m = 1$

Dynamic 1RSB phase

$\mu \in [\mu_c, \mu_d] \Rightarrow$ Exponential number of clusters contributing to the Gibbs measure. $\Sigma(m = 1) > 0$
RS predictions for thermodynamic quantities are correct

Condensate 1RSB phase

$\mu < \mu_c \Rightarrow$ Only a sub-exponential number of clusters contributes to the Gibbs measure. $\Sigma(m = 1) < 0$
The thermodynamic properties are the ones of the clusters selected by the value of m_s s.t. $\Sigma(m_s(\mu)) = 0$

Pattern in CSP

As the Constraint Satisfaction Problem becomes harder (i.e. looking for smaller and smaller θ), the space of the solutions can be pictorially represented as follows:

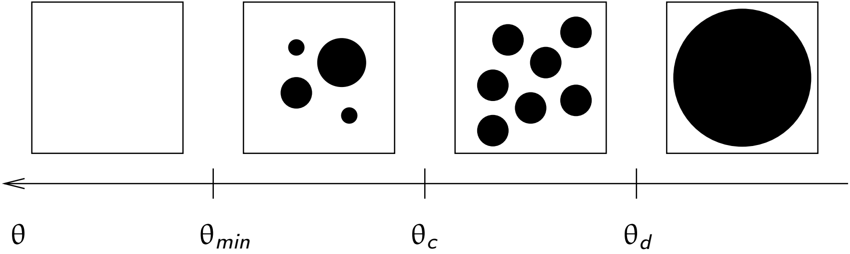


Table of Contents

Energetic 1RSB

Simplified version of 1RSB when $\varepsilon = +\infty$, $\mu \to -\infty$
$\Sigma(s,\theta)$: number of clusters with e^{Ns} percolating configurations with $\theta\, N$ seeds (whose free entropy is $\phi = \mu\,\theta + s$)

$$\Phi(m) = \sup_{\theta,s}[\Sigma(s,\theta) + m(\mu\,\theta + s)]$$

Energetic 1RSB

If $m \to 0$ and $\mu \to -\infty$ with a finite value of $y \equiv -\mu m$:

$$\begin{cases} \Phi_e(y) = \sup_\theta[\Sigma_e(\theta) - y\theta] \\ \Sigma_e(\theta) = \sup_s \Sigma(s,\theta) \end{cases} \Rightarrow \begin{cases} \Sigma_e(\theta(y)) = \Phi_e(y) + y\theta(y) \\ \theta(y) = -\Phi_e'(y) \end{cases}$$

$$\theta_{gs,1RSB} = \theta(y_s) \text{ with } y_s \text{ s.t. } \Sigma_e(\theta(y_s)) = 0$$

The Warning Propagation equations

With the previous assumptions, the fields h can take only $2T+1$ values: A_t for $t \in [0, T-1]$ and B_t for $t \in [0, T]$

Definitions

The message $h_{i\to j}$ is a **warning** sent from node i to one of its neighbours j.
$G_{i\setminus j}$ is the subtree rooted at i excluding j

Intuitive interpretation

$h_{i\to j} = B_t \Rightarrow t_i = t$ and $G_{i\setminus j}$ activates before T with $\sigma_j^T = 0$.
$h_{i\to j} = B_0 \Rightarrow i$ is a seed
$h_{i\to j} = A_t \Rightarrow G_{i\setminus j}$ activates before T only if $\sigma_j^t = 1$

The combination among these warnings gives the relations
$h = g(h_1, ..., h_k)$

Table of Contents

The large T limit

- $T \to \infty$: **original influence maximisation problem**
- The limit can be solved analytically
- $k = l$ and $k > l$ are qualitatively different cases

Results

$k = l$

Already known lower bound: $\theta_{min}(k,k) \geq \frac{\theta_r}{2} = \frac{k-1}{2k}$

- For $k = l = 2 \Rightarrow$ Saturation of the bound: $\theta_{min}(2,2) = \frac{\theta_r}{2} = \frac{1}{4}$
- For $k = l = 3 \Rightarrow$ Saturation of the bound: $\theta_{min}(3,3) = \frac{\theta_r}{2} = \frac{1}{3}$
- For $k = l \geq 4 \Rightarrow$ Unsaturation of the bound: 1RSB predictions have been obtained

$k > l$

Already known lower bound: $\theta_{min}(k,l) \geq \frac{2l-k-1}{2l}$

- For $k = 4, l = 3 \Rightarrow$ Saturation of the bound: $\theta_{min}(4,3) = \frac{1}{6}$
- For $k = 5, l = 4 \Rightarrow$ Saturation of the bound: $\theta_{min}(5,4) = \frac{1}{4}$
- For larger values $\Rightarrow$ Unsaturation of the bound: 1RSB predictions have been obtained

Table of Contents

The single sample analysis

- Explicitely define a $(k+1)$ RRG (for different k) (e.g. with $N = 10000$)
- Run different algorithms
  - Start without seeds
  - Add the seeds one by one according to some rule
  - Stop when a percolating configuration is found
- Compare the minimal densities of the percolating configurations found

Stability of the results

Both for different instances and for different runs on the same graph

The greedy algorithm at finite T

Starting point

$\sigma_i^0 = 0 \quad \forall i$

Iteration

- Simulate the contagion adding one extra seed
- Set as seed the node improving the most $\sum_i \sigma_i^T$

Ending point

Stop when a percolating $\underline{\sigma}$ is found

Greedy algorithm at $T = \infty$

Generalisation

For monotonicity and finite size, one reaches $\underline{\sigma}^\infty = \lim_{T\to\infty} \underline{\sigma}^T$ in finite time.

Computational simplification

When choosing an extra-seed, no need to restart from $\underline{\sigma}^0$: one can start from the $\underline{\sigma}^\infty$ of the previous iteration

The MP 1RSB algorithm

Energetic 1RSB potential

$$\Phi_e(y) = -y + \frac{1}{N}\sum_{i=1}^{N} \log(\mathcal{Z}_{site}(\{P_{j\to i}\}_{j\in\partial i})) - \frac{1}{N}\sum_{\langle i,j\rangle\in E} \log(\mathcal{Z}_{edge}(P_{i\to j}, P_{j\to i}))$$

Knowing that $\theta(y) = -\Phi'(y)$, one can define a **score** as the contribution of node i to the overall $\theta(y)$:

Score of a node i

$$S(i) = 1 - \partial_y \log \mathcal{Z}_{site}(\{P_{j\to i}\}_{j\in\partial i}) + \frac{1}{2}\sum_{\langle i,j\rangle\in E} \partial_y \log(\mathcal{Z}_{edge}(P_{i\to j}, P_{j\to i}))$$

The decimation strategy

Starting point

$\sigma_i^0 = 0 \quad \forall i$

Iteration

- Run the MP iterations till the convergence of the messages
- Calculate $S(i)$ for each node not yet fixed to seed
- Fix to seed the one with the largest score
- Fix its out-going messages to δ_{q_0} and stop updating them

Ending point

Stop when a percolating $\underline{\sigma}$ is found

Comparison of the results

N=10000, K=L=2

	Greedy	MP-1RSB	θ_{min}
T=1	0.4821 ± 0.0005	0.42589 ± 0.00004	0.424
T=3	0.3350 ± 0.0003	0.29112 ± 0.00003	0.289
T=5	0.2958 ± 0.0002	0.26313 ± 0.00002	0.262
T=∞	0.25013 ± 0.00001	?	0.25

N=10000, K=3, L=2

	Greedy	MP-1RSB	θ_{min}
T=1	0.4264 ± 0.0004	0.36650 ± 0.00007	0.363
T=3	0.2328 ± 0.0002	0.18533 ± 0.00004	0.182
T=∞	0.0709 ± 0.0001	?	0.0463

Non-convergence of MP

For large enough T, the MP-1RSB iterations do not converge anymore.
In particular, for $k = 3, l = 2$ iterations converge just up to T=3.
The reasons are still to be fully understood

Results for larger T

At each step of the decimation, update the messages a fixed number of times

Comparison in the non-convergence region

- For small enough T, still good results of MP-1RSB
- As T increases, $\theta_{MP-1RSB} - \theta_{min}$ increases
- MP-1RSB is anyway still better than greedy algorithm

N=10000, K=3, L=2

	Greedy	MP-1RSB	θ_{min}
T=4	0.1975 ± 0.0002	0.15610 ± 0.00003	0.1517
T=5	0.1742 ± 0.0002	0.14200 ± 0.00005	0.1320
T=7	0.1442 ± 0.0002	0.12697 ± 0.00007	0.1083

Table of Contents

Open points

Convergence issues

Possible convergence of 1RSB for larger T?

Generalisations

Go beyond $k_i = k, l_i = l \quad \forall i$. In particular:

- Fluctuating connectivities (even with constant threshold)
- Fluctuacing activation thresholds (even on regular graphs)

Finite ε

Interesting problems: for example maximum possible spread with a fixed number of seeds

Possible applications

Applications to real-world networks