Università degli Studi di Padova
Department of Mathematics "Tullio Levi-Civita"
Master Degree in Data Science

Linear Threshold Rank on Random Social Networks

Supervisor: Prof. Marco Formentin
Department of Mathematics
Co-Supervisor: Prof. María José Serna Iglesias
Department of Computer Science, UPC

Student: Laura Iacovissi
N. 1205451

Academic Year 2019-2020

Contents

Abstract

The Linear Threshold Rank (LTR) is a centrality measure based on the Linear Threshold Model for influence spread. In this thesis we study the LTR on two random graph models: the Erdös–Rényi Graphs and the Random Geometric Graphs. The main focus is on the impact the threshold definition have on the algorithm output. Two kind of deterministic thresholds are considered: a natural one, the percentage of neighbours that must be active; an approximation of the first one, that replaces the number of neighbours with the maximum number of connections a node can have.

The experiments show similar behaviors for the two thresholds on both models (directed and undirected), even if the approximated version resulted faster. We notice some interesting properties with phase transitions. It is also observed that, in the connected regime and with an increasing percentage value in the threshold definition, the metric goes from being maximum to count only the nodes in the initial activation set. In correspondence of this change in the metric value, the maximum levels have a peak and the ranking assumes values in a larger range.

After the experiments discussion, some theoretical results are proved only for the Erdös–Rényi model.

Introduction

In a wide range of real-world cases, data can be organized as a network. Understanding which elements of the network are, according to some definition, more relevant in the structure is often a useful way to understand data. Centrality measures aim to determine how important is an element of the network (node, vertex, actor) within the structure itself. Some classical examples of centrality measures are the *degree metric*, which judges a node important counting how many connections it has, or the *betweenness metric*, which considers the number of shortest paths passing through the current vertex. In particular, when dealing with interaction networks like social networks, networks of particles or communication networks, the definition of centrality can be based on how an element can *influence* the others. In this work, this whole class of networks will be generally called *social networks*. They are formally represented as graphs where the interactions are the edges. Each edge can be associated with a numerical measure of the interaction strength, called weight.[1]

An influence expansion model describes the ways in which actors influence each other through their interactions in a social network. The most famous and studied ones are the Linear Threshold Model and the Independent Cascade Model [30]: the first one is based on some ideas of collective behavior [28, 44], while the second one was proposed in marketing contexts [26].

Recently, in [42], it has been introduced a centrality measure based on the Linear Threshold Model, called the Linear Threshold Rank (LTR). This measure evaluates node importance depending on how many nodes it can influence during the expansion process, assuming it able to convince his immediate neighbours. Given this initial effort, even nodes with a low number of connections can result highly central thanks to their neighbourhood.

In order to be able to perform the influence expansion on a graph, we need an additional information about the *resistance* of each node to be influenced. This quantity is usually not known a priori when data are collected, so finding a way to define meaningfully the resistance values is a key problem. To this end, different

[1] From now on, the words graph and network, as well as node, vertex, actor and edge, link, interaction will be used as interchangeable.

studies have been performed on benchmark networks [42, 43, 45, 23, 16] with different resistance assignments. However, in this work, it is analysed the behavior of the Linear Threshold Rank for the first time on random network models, defining the resistances in a deterministic way.

The random graph models used for this purpose are the Erdös–Rényi Graphs and the Random Geometric Graphs. The former has been chosen because it is the simplest definition possible for a random network, so that a preliminary study of LTR can be performed. The latter has connectivity regimes similar to the Erdös–Rényi Graphs, so that the results can be compared with the ones on the first model in a more immediate way. In addition, the Random Geometric Graphs have been used in real networks modeling [5] and in continuum percolation [2, 3, 41] - almost exclusively in two and three dimensions.

The thesis is structured in the following way:

- Chapter 1: an overview on information diffusion models is given. The Linear Threshold Model is formally introduced together with the Linear Threshold Rank. The main studies on the centrality measure are reviewed;
- Chapter 2: the random graph models used in the work are defined and briefly described, in order to give to the reader all the tools needed to fully understand the results;
- Chapter 3: the concepts of Random Graph and Influence Graph are mixed, hence the needed definitions and notations are introduced;
- Chapter 4: the general framework of the experiments is explained. Each numerical simulation is described in detail, organizing it in different phases. For each phase, the results are presented and discussed;
- Chapter 5: some of the properties arose from the experiments on the Erdös–Rényi model are formally proved;
- Chapter 6: the conclusion and future work of the thesis are discussed.

Chapter 1

Influence Expansion

1.1 Related Work on Influence

Social influence analysis (SIA) is becoming an important research field in social networks. SIA mainly studies how to model the influence diffusion process in networks, and how to propose an efficient method to identify a group of target nodes in a network [9]. Studied questions include: who influences whom; who is influenced; who are the most influential users.

SIA models have been widely studied in the literature. In [33], they are divided into two main categories: microscopic and macroscopic models.

Here we recall the definition of graph, in order to fix the notation.

Definition 1.1. A **graph** (**digraph**) G is an ordered pair (V, E) of sets, where

$$V : \text{set of nodes}$$
$$E = \{\{i, j\} \mid i, j \in V\} : \text{set of unordered edges}$$
$$or$$
$$E = \{(i, j) \mid i, j \in V\} : \text{set of ordered edges}$$

Microscopic Models

Microscopic models focus on the role of individuals interactions and examine the structure of the influence process. The most famous and studied influence analysis models in this category are the Linear Threshold Model (LTM) and the Independent Cascade Model (ICM) [30]. Since the LTM will be the basis of this thesis, here we will briefly describe only the ICM and some alternative models.

ICM. Consider a graph $G = (V, E)$, a seed set $S \subset V$ and $S_t \subset V, t \geq 0$ the set of nodes that are activated at step t. At step $t + 1$, every node i can activate its out-neighbour j with an independent probability p_{ij}. The process ends when no node can be activated. Note that a node has only one chance to activate

its out-neighbors after it has been activated, and the node cannot exit from the "activated" state.

Alternative models. Some models are different from the ICM or LTM models and their variations, and have solved information influence diffusion from a new point of view.

Lin et al. [34] proposed a data-driven model to maximize the expected influence in the long run. However, this model needs large amounts of data, and the accuracy of its results requires further improvement.

Golnari et al. [27] proposed a heat conduction (HC) model. It considers a non-progressive propagation process, and is completely different from the previous ICM or LTM models, which only consider the progressive propagation process. In the HC model, the influence cascade is initiated from a set of seeds and arbitrary values for other nodes.

Wang et al. [46] studied emotion influence in large-scale image social networks, and proposed an emotion influence model. They designed a factor graph model to infer emotion influence from images in social networks.

Gao [22] proposed a read-write (RW) model to describe the detailed processes of opinion forming, influence, and diffusion. However, there are three main issues that this model needs to consider further: the many parameters of the model that must be inferred, the proper collection of datasets about opinion influence and diffusion, and the evaluation metrics that are suitable for this task.

Macroscopic Models

Macroscopic models consider all users to have the same attraction to information, the same transmission probability, and identical influential power. However, since macroscopic models do not take individuals into account, the accuracy is lower. To improve such models, the differences between individuals should be took into account. Epidemic models are the most common models that are used to study social influence from a macroscopic perspective. These models were mainly developed to model epidemiological processes. However, they neglect the topological characteristics of social networks.

Daley–Kendall model. Daley and Kendall [11] analysed the similarity between the diffusion of an infectious disease and the dissemination of a piece of information, and proposed the classic Daley–Kendall model. Since then, researchers have improved these epidemic models in general to overcome their weaknesses.

SI and variations. A standard and basic model belonging to the Epidemics Models class is the Suceptible Infected (SI) one, proposed in [38]. The model assumes that the total number of people, equal to N, is divided into two categories: S (susceptible) and I (infected). At time t, $s(t)$ represents the susceptible proportion of the total population, $i(t)$ represents the infected proportion and $s(t) + i(t) = 1$.

The λ parameter represents the daily contact rate, i.e. the proportion of the susceptible users infected by infected users in the total population.

Given these definitions, we have that there will be $N\lambda s(t)i(t)$ susceptible users infected per day. Assuming at time $t = 0$ the proportion of patients is i_0 and noticing that $\frac{di}{dt} = \lambda si = \lambda(1-i)i$ we have

$$\begin{cases} \frac{di(t)}{dt} & = \lambda(1-i(t))i(t) \\ i(0) & = i_0. \end{cases} \tag{1.1}$$

This model can be expanded adding different kind of node labels. There are various model developed specifically for social networks, for example SEIR (Suceptible Exposed Infected Removed) model [47], S-SEIR (Single layer-SEIR) [49], SCIR (Suceptible Contacted Infected Removed) model [15], irSIR (infection recovery SIR) model [8], FSIR (Fractional SIR) model [19] and ESIS (Emotional Suceptible Infected Suceptible) model [48].

1.2 Linear Threshold Model

The Linear Threshold Model is a deterministic modelization of the influence spread phenomena across a network: it describes step-by-step the process that leads a network member (*actor*) to influence the other members [30]. One of the key features of this model is that it is *progressive*: an actor that receives influence from another actor for the first time turns its state from inactive to active in an irreversible way.

To formalize the LTM dynamics, it is necessary to give the definition of influence graph [42].

This definition given in Def. 1.1 can be expanded in order to allow the structure carry more information about the object it is modeling. In particular, when dealing with interaction networks, like for example social networks, it is useful to define the influence graphs.

Definition 1.1. An **influence graph** is a triple (G, w, f), where:

$G = (V, E)$ is a graph (digraph);
$w : E \to \mathbb{R}$ is a function on edges, called weight function;
$f : V \to \mathbb{R}$ is the labeling function, which assigns to each node its resistance to be influenced. It also known as activation threshold.

The above definition is clearly valid either for directed or undirected graph. For undirected graphs, an associated directed graph can be obtained by considering every edge in both directions. Hence, in this section the notation will assume that the graph G is a digraph.

Figure 1.1: Chain with $n = 4$.

Given an influence graph (G, w, f) and an initial set $X \subseteq V$ of active nodes, consider the following iterative activation process. Let $F_t(X) \in V$ be the set of **active nodes** at iteration t, with $F_0(X) = X$. At each step $t > 0$ the LTM prescribes that new nodes may be added if the following condition is satisfied by each of them. Given $F_{t-1}(X)$, $x \in F_t(X)$ if and only if

$$\sum_{y \in F_{t-1}(X)} w(y, x) \geq f(x) \tag{1.2}$$

where $w(y, x)$ is the weight associated to the x's incoming edge $\{y, x\} \in E$ and $f(x)$ is the resistance associated to $x \in E$ [30]. We put $w(y, x) = 0$ when $\{y, x\} \notin E$, i.e. when weight is not defined. The process stops when an iteration cannot activate any new node.

The influence process is here formally defined, as done in [42]:

Definition 1.2. Let (G, w, f) be an influence graph where $G = (V, E)$. The **spread of influence** of $X \subseteq V$ is

$$F(X) = \bigcup_{t \geq 0} F_t(X).$$

At each step the set of active nodes is updated in the following way:

$$F_t(X) = F_{t-1}(X) \cup \{\text{nodes activated at step } t\}. \tag{1.3}$$

We have that $\{F_t(X)\}_{t \geq 0}$ is a strictly monotone, increasing sequence of sets.

Notice that the number of time steps t is at most equal to $n - 1$ when the number of vertices in the graph is n, as if the process does not stop it incorporates at least one vertex. The number of steps is maximized, for example, when the graph is a chain of length n (example in Figure 1.1): under this condition the spread of influence would increase only of one node at each step. In general, as shown in the example in Fig. 1.2, it is true that

$$|F_t(X)| - |F_{t-1}(X)| \geq 1 \tag{1.4}$$

before the process reaches the stopping condition

$$F_t(X) = F_{t-1}(X). \tag{1.5}$$

Our next result shows that, given X, the set $F(X)$ can be computed efficiently.

Algorithm 1 LTM Influence expansion algorithm.

```
1: Given (G, w, f): influence graph, X initial set
2: Initialize total = 0, Q empty queue (FIFO)
3: for v ∈ G.V \ X do
4:     v.active = FALSE
5:     v.influence = 0
6:     v.level = −1
7: for v ∈ X do
8:     v.active = TRUE
9:     v.influence = 0
10:    v.level = −1
11:    ENQUEUE(Q, v)
12: total = length(Q)
13: while Q not empty do
14:    v = DEQUEUE(Q)
15:    for u ∈ G.neighbours[v] do
16:        if not u.active then
17:            u.influence = u.influence + w(u, v)
18:            if u.influence ≥ f(u) then
19:                u.active = TRUE
20:                total = total + 1
21:                ENQUEUE(Q, u)
22:            u.level = v.level + 1
   return total
```

Theorem 1.3. *Let (G, w, f) be an influence graph and X a subset of nodes of G. There is an algorithm that, with input (G, w, f) and X, computes the spread of influence $F(X)$ with time complexity of $O(|E| + |V|)$ when G is represented with adjacency list, in $O(|V|^2)$ when represented with adjacency matrix.*

Proof. In order to implement efficiently the computation the $F(X)$ defined in Def. 1.2 we use a different but equivalent approach, see Algorithm 1.
The code has three main steps: first, it labels the vertices as *active* or not (`v.active`); then, for each vertex it is computed the currently received *influence* (`v.influence`), initially set to 0; finally, a queue Q (FIFO - First In First Out data structure) is used as a temporary placement for activated nodes whose influence have not been expanded yet.
Initially the vertices in Q are only the initial active nodes in X. While the queue is not empty, The algorithm will process the vertices as wrote in lines 13-22: the first active node v in the queue is extracted and the neighbours receive its influence. The $w(u, v)$ influence exerted on u, a neighbour of v, is added to the influence

parameter. If the level of influence reaches or exceed the threshold, v is deemed as active and added to the back of the queue.

Now we can prove by induction that the algorithm is correct:

1. Observe that the queue respects the order of activation of the vertices, since it is a FIFO structure. Initially $F_0(X) = X$ is activated, and placed on the queue. When the algorithm finishes processing those vertices in $F_0(X)$, all the vertices in $F_1(X) \setminus F_0(X)$ have been placed at the end of the queue and the ones in $F_0(X)$ deleted from it;

2. Iteratively, when the algorithm finishes treating the last vertex in $F_t(X)$, all and only the vertices in $F_t(X) \setminus F_{t-1}(X)$ are in the queue. Therefore, when the queue is empty, no more vertices can be activated, and the set of active vertices coincides with $F(X)$.

For the complexity part, observe that:

1. lines 3-11: we have two loops that initialize the parameters. The cost is obviously $\Theta(|V|)$;

2. line 12: constant cost $\Theta(1)$;

3. lines 13-22: the `while` loop stops in at most in |V| iterations, since once we extract a vertex from the queue it never enters it again; for each vertex, we update the information of its neighbours (constant cost if efficiently implemented with a doubly linked list). The cost of the loop depends on the computational cost of the not active neighbours retrieval for each node.
   When we have an adjacency list, in the worst case - i.e. when all neighbours are not active - this cost is $O(|neigh(i)|)$ with $\{neigh(i)\}_{i \in V}$ disjointed sets; when an adjacency matrix is $O(|V|)$.
   Hence, for the first case the total cost for the list exploration is the sum of $O(|neigh(i)|)$ over all the nodes $i \in V$, equal to $O(|E|)$. Since the access to the lists is $O(|V|)$, then the cost of the whole lines is $O(|V| + |E|)$.
   In the second case, the two nested loops have same maximum number of iterations $O(|V|)$, so the total cost is $O(|V|^2)$.

Hence, the claim follows. □

In Algorithm 1, we have added some counters to keep track of relevant parameters that we will use to analyse the properties of the process.

Definition 1.4. Let (G, w, f) be an influence graph where $G = (V, E)$. A node is said to be **visited** at step t if the condition in equation (1.2) has been checked at this step.

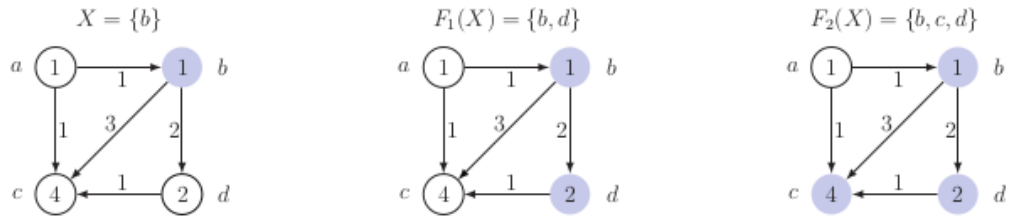


Figure 1.2: A simple example is proposed in order to clarify the mechanism of the LTM. At level $t = 0$ only the initial set $X = \{b\}$ is activated; at $t = 1$ the neighbours of b receive the its influence and only d has a resistance value low enough to be activated, the other neighbour c is just visited (its level variable is updated to 1 but it is not active); at level $t = 2$ the node c is reached by the influence of b and d so it is activated. The only inactive node at $t = 2$ is a: it cannot be reached by any influence, since it has no incoming connection. Its level will remain -1 and in this case the maximum level and the spread level are the same.

Definition 1.5. Let (G, w, f) be an influence graph where $G = (V, E)$. The set of **visited nodes** starting from the initial activation set X is

$$visited(X) = \bigcup_{t \geq 0} visited_t(X).$$

Notice that the sequence $\{visited_t(X)\}_{t \geq 0}$ is monotone but not strictly, since nodes can be visited more than one time during the influence spread process.

Definition 1.6. Let (G, w, f) be an influence graph where $G = (V, E)$.
The **level** of a node $x \in V$ starting from the initial activation set X is the value of t at which it has been visited for the last time. If x is never visited, the level is assumed to be -1.

Looking at the Algorithm 1, we can see that the update of the level attribute is done outside the `if` condition (line 22). This explains why the definition Def. 1.6 talks about visited nodes, not active ones. Of course, when a node is activated (enters the $F_t(X)$ set for some level t) its level cannot be updated anymore.

Definition 1.7. Let (G, w, f) be an influence graph where $G = (V, E)$.
The **spread level** k_X from the initial activation set X is defined as

$$k_X = \min\{t \geq 0 \mid F_{t-1}(X) = F_t(X)\}$$

and it is the maximum level an active node can reach starting from X.

Algorithm 2 BFS algorithm.

```
1: Given G: graph, x ∈ V initial node
2: Initialize component_size = 1, Q empty queue (FIFO)
3: for v ∈ G.V \ {x} do
4:     v.active = FALSE
5:     v.distance = −1
6: x.active = TRUE
7: x.distance = 0
8: while Q not empty do
9:     v = DEQUEUE(Q)
10:    for u ∈ G.neighbours[v] do
11:        if not u.active then
12:            u.active = TRUE
13:            u.distance = v.distance + 1
14:            component_size = component_size + 1
15:            ENQUEUE(Q, u)
   return component_size
```

Because of the observation after Def. 1.2, we have $k_X < |V|$.

Definition 1.8. Let (G, w, f) be an influence graph where $G = (V, E)$. The **maximum level** from the initial activation set X is defined as

$$max\_level_X = \min\{t \geq 0 \mid visited_{t-1}(X) = visited_t(X)\}$$

and it is the maximum level a node can reach starting from X.

Remark 1.9. By definition of all the considered quantities, we can easily see that the following equivalence holds

$$max\_level_X - k_X = \begin{cases} 0 & \text{if } F(X) = visited(X) \\ 1 & \text{if } F(X) \neq visited(X) \end{cases}$$

since the only possible case for $F(X) \neq visited(X)$ is $F(X) \subset visited(X)$, when there are inactive node connected with nodes in $F(X)$ but not enough influenced.

The level, as defined in Def. 1.7, is the t at which a node is activated. It is important to underline that, differently from a graph traversal, more than one node can be added to the $F_t(X)$ set for each t. This means that the level is not interpretable as a time, since in order to add all the activable node to the current active set more than one iteration of the `while` cycle can be performed (lines 13-22 of Alg. 1). Depending on how the influence expansion is implemented, it could

be connected to the notion of *distance* from the initial activation set X. In the algorithm reported in Alg. 1 the spread level variable is actually correlated to the distance a node have from the set X, since it cannot assume a smaller value. This property can be explained by comparing this algorithm with the *Breadth-First Search* (BFS) one (Alg. 2): the two algorithms explore the graph in a similar way (the adjacent of a node are visited and a FIFO queue is used to store the influenced nodes, lines 13-21 of Alg. 1, lines 6-12 of Alg. 2); on the other hand, the LTM add a constrain for the visited node to be activated, which radically change the order in which the nodes are added to the queue. The set of active nodes returned by the LTM algorithm can be interpreted as a *influenced component*, given the similar way the variables *total* (LTM) and *component_size* (BFS) are updated in the two algorithms.

1.3 Linear Threshold Rank

Starting from the LTM for influence expansion, a new centrality measure has been proposed in [42]. It assigns to each node a centrality value in the $[0, 1]$ interval by considering the number of actors of the graphs it is able to affect with its influence.

Definition 1.1. Let (G, w, f) be an influence graph, with $G = (V, E)$ and $x \in V$ an actor. The **Linear Threshold Rank** of x, denoted by $LTR(x)$, is given by

$$LTR(x) = \frac{|F(\{x\} \cup neigh(x))|}{|V|}$$

where $neigh(x) = succ(x) \cup pred(x) = \{y \in V \mid (x, y) \in E \vee (y, x) \in E\}$: set of all the nodes with a connection with x.

The choice of the $neigh(x)$ set definition is crucial to determine which kind of property our ranking is going to capture. The consequences of this choice have been studied in [45, 23, 43, 16], together with modifications of the LTR. Here some of the most relevant results are summarized. Notice that all the studies reported below have performed analyses on real data (benchmark networks, for example the arXiv network[1], the Higgs network[2] and the Wikipedia voting network[3]), while this work will focus on the LTR features on some *graph models*.

In [45] the LTM is studied with a threshold of *simple majority* (the labeling function defined as $f(x) = 0.5 \cdot \sum_{y \in neigh(x)} w_{yx}$). The definition of the neighbourhood set has been stated as the set of successors of a node (*Forward Linear Threshold Ranking*, FLTR): this has been justified with a theoretical example, which have

[1] `https://arxiv.org/archive/gr-qc`

[2] `http://snap.stanford.edu/data/higgs-twitter.html`, reference paper: [12]

[3] `http://snap.stanford.edu/data/wiki-Vote.html`, reference paper: [32]

shown that other definitions might have resulted in height of the rankings which does not reflect the real cases. The example is the following: consider a Twitter network where nodes correspond to Twitter accounts. In this framework, edge (i, j) belongs to the graph when j is a follower of i. Imagine that node i decides to follow a celebrity with many followers: that should not imply an increase in the influence rank of node i, since they are not able to spread their influence through this new connection. However, the LTR defined with $neigh(i) = succ(i) \cup pred(i)$ increases in this case: this celebrity-node becomes part of the initial activation set of node i and therefore leads to a higher LTR. This setting obviously lead to a misinterpretation of the influence spreading process that happens in real-world cases.

In addition, two more general rankings based on LTM have been studied: the *Discounted Linear Threshold Rank* (DLTR) and the *Fading Linear Threshold Rank* (FALTR). These last two measures of centrality are different implementations of the same idea: evaluate the relevance of a node not only considering the number of actors it is able to influence, but also how much time steps it needs to achieve the actors activation. The former method have given results similar to the LTR's ones, with the relevant difference of having less nodes with the same metric value. This means that, with properly set parameters, the DLTR can better distinguish the role of a node in the considered network. The latter model, FALTR, had the same advantages of DLTR and the disadvantage of having a higher computational cost.

In [23] different aspects of the LTR have been taken into consideration: the study of the discounted version of the LTR has been expanded; the role of the neighbourhood has been analysed for DLTR and for the standard LTR; a *Bounded Linear Threshold Rank* (BOLTR) has been defined and compared with LTR. Again, the threshold used is the *simple majority*.

Regarding the DLTR, the conclusions are similar to the ones produced in [45]: the results and the execution costs are almost the same of the LTR, with the advantage of having more differentiation on the metric outputs.

For the analysis of the neighbourhood role, it has been defined in three different ways: $succ(x) \cup pred(x)$ (LTR), only $succ(x)$ (*Forward* flavoured, FLTR), only $pred(x)$ (*Backward* flavoured, BLTR). The centrality measures produced by LTR and BLTR are highly similar and seem to associate to each node a false potential of influence, above the one that has in a real world networks. The FLTR gave different results, more in line with the real world cases.

The last considered aspect of the LTR was the BOLTR, defined depending on a parameter representing the maximum spread level a node is allowed to reach. This ranking showed to be almost identical to the standard LTR even with low values of the bound: this shows that the LTR is, on these graphs, not able to

influence nodes at high distance from the initial set.

In [43] it is defined a generalization of the LTR based on a flexibilization at the level of neighbourhood considered for the initial activation: the Linear Threshold Rank of x at level l, where l represents the maximum distance a node can have from the seed of the influence expansion in order to be included in the neighbourhood. This metric, applied in combination with different definition of the labeling function, for the first time defined not only as a simple majority, have shown that it can be obtained even more distinguishable ranking values by using higher levels of neighbours. The threshold used are: minimum influence ($f(x) = 1$), maximum influence ($f(x) = \sum_{y \in neigh(x)} w_{yx}$), simple majority, random.

In [16] the main focus was on the labeling function definition for the FLTR. Three possible definitions have been considered: with the percentage of required active nodes constant on all nodes, randomly generated and generated by another measure of centrality. Note that the weights of the graphs have been normalized s.t. the sum of the incoming weights for a specific node is always less than one and the thresholds can be picked in the $[0, 1]$ interval.

The constant labeling function case showed that an inflection point in the $[0.2, 0.5]$ interval exists, where the number of influenced actors decreases quickly.

The random threshold assignment showed to allow the nodes to have high capability of influence, even if the interval from which the labels are sampled is $[0.5, 1]$.

For the third kind of f initialization, the centrality measures used as threshold were *Betweenness* [20], *ICR* [29], *PageRank* [37] and the same FLTR. The last two gave really similar results (compared through different correlation coefficients). In general, this assignments based on other metrics seem to be more restrictive with respect the other two tried in this work, i.e. low values of ranking are generated.

In this work we will focus on only one version of the LTR, the following one:

Definition 1.2. Let (G, w, f) be an influence graph, with $G = (V, E)$ and $x \in V$ an actor. The **Forward Linear Threshold Rank** of x, denoted by $FLTR(x)$, is given by

$$FLTR(x) = \frac{|F(\{x\} \cup succ(x))|}{|V|}$$

where $succ(x) = \{y \in V \mid (x, y) \in E\}$ is the set of all the x successors.

In the undirected graph case it is equivalent to the LTR.

Chapter 2

Random Graphs

Depending on the definitions of the sets V and E (see Def. (1.1)) a graph can be deterministic, random, dynamic (it evolves in time) or assume any kind of flavour that best suits the data we want to represent.

In this section we will focus on two well known definition of random networks, describing some of their main properties: the Erdös–Rényi and the random geometric graphs. With the term *random* we refer to graphs with a probability distribution defined over them.

2.1 General definitions

In order to fix the notation that will be used in the section, some basic definitions on graphs are recalled.

Definition 2.1. Given a graph $G = (V, E)$, the **degree** $d(i)$ of a vertex $i \in V$ is defined to be the number of vertices connected with i by an edge. More formally:

$$d(i) = \sum_{j \in V \setminus \{i\}} \mathbb{1}(\{i, j\} \in E) \, .$$

In the case of directed graph, the degree can be defined as in-degree $d^{IN}(i)$ considering only the predecessors, out-degree $d^{OUT}(i)$ only with successors or the total degree $d(i) = d^{IN}(i) + d^{OUT}(i)$.

Definition 2.2. Given a graph (digraph) $G = (V, E)$, a node $i \in V$ is said to be **isolated** if $d(i) = 0$.

Definition 2.3. Given a graph $G = (V, E)$, a **connected component** $C = C(G)$ is a subgraph of G s.t. $C = (V', E')$ with $V' \subset V$, $E' \subset E$ where any two vertices in V' are connected to each other through a path and vertices of V' are not connected to vertices of $V \setminus V'$. When the size of $C(G)$ is of the order of the entire graph, then we call it a **giant component** (GC).

Definition 2.4. Let $G(n, \theta_n)$ be a random graph of size n depending on some parameters $\theta_n = (\theta_n^1, \dots, \theta_n^d) \in [0,1]^d,\ d \geq 1$. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence in the real interval $[0,1]$. A property $\mathcal{P}$ is said to have a **sharp phase transition** with respect the to converging sequence $(a_n)_{n\in\mathbb{N}}$ and the parameter θ_n^i if

$$\lim_{n\to+\infty} \mathbb{P}(\mathcal{P} \text{ holds for } G(n,\theta_n)) = \begin{cases} 0 \text{ if } \theta_n^i \geq c \cdot a_n,\ c > 1 \\ 1 \text{ if } \theta_n^i \leq c \cdot a_n,\ c < 1 \end{cases}$$

or

$$\lim_{n\to+\infty} \mathbb{P}(\mathcal{P} \text{ holds for } G(n,\theta_n)) = \begin{cases} 0 \text{ if } \theta_n^i \leq c \cdot a_n,\ c < 1 \\ 1 \text{ if } \theta_n^i \geq c \cdot a_n,\ c > 1 \end{cases}$$

when all the other components of θ_n are fixed.

2.2 Erdös–Rényi Graphs

We now extend the definition of graph given in Section 1.2 introducing a simple random component, which makes it a *random graph*. The following definition of Erdös–Rényi graph (ERG) or binomial graph has been introduced in [17, 24, 18]; this model associates to the edges a probability of appearance in the simplest way possible.

Definition 2.1. An **Erdös–Rényi graph** $G(n,p)$ is a graph $G = (V,E)$ with $n = |V|$ and where each possible edge has probability p of existing, i.e.

$$\begin{aligned} &E = \{\{i,j\} : \xi_{i,j} = 1,\ i,j \in V,\ i \neq j\} \\ &\xi_{i,j} \sim Be(p)\ i.i.d. \end{aligned}$$

where $Be(p)$: Bernoulli distribution of parameter p.

It can be defined in a directed or undirected flavour. The notation used above is for the undirected case. i.e. $(i,j) = (j,i)$, while in the directed case we have:

$$\begin{aligned} &E = \{(i,j) : \xi_{i,j} = 1,\ i,j \in V,\ i \neq j\} \\ &\xi_{i,j} \sim Be(p)\ i.i.d. \end{aligned}$$

where $(i,j) \neq (j,i)$ and $\xi_{i,j} \neq \xi_{j,i}$.

Now we can describe some of the main properties of an Erdös–Rényi graph.

Expected degree. This definition allow us to easily compute the *expected degree*. Consider the discrete random variable $d(i)$ (or $d^{IN}(i)$, $d^{OUT}(i)$ in the directed case) representing the degree value of the $i \in V$ node of a binomial graph. Given the Def. 2.1, we have that:

$$\mathbb{P}(d(i) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}. \tag{2.1}$$

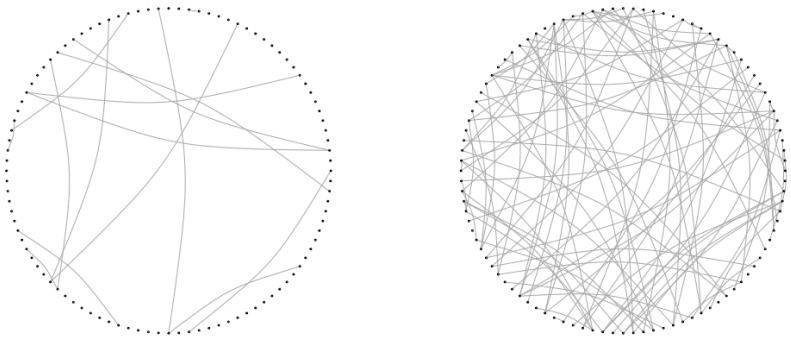


Figure 2.1: Examples of $G(n,p)$ graphs with $n = 100$. On the left, $p = \frac{1}{300}$ (isolated regime); on the right, $p = \frac{1}{50}$ (GC regime).

In other words, $d(i) \sim B(n-1,p)$: Binomial random variable, so its expected value and variance are:

$$\begin{aligned} \mathbb{E}(d(i)) &= (n-1)p \ ; \\ \mathrm{Var}(d(i)) &= (n-1)(1-p)p \ . \end{aligned} \tag{2.2}$$

Connectivity regimes. Interesting properties of this class of graphs are the asymptotic ones, usually described through sharp phase transitions with respect to the probability parameter (see Def. 2.4). In literature, the results are only on the *undirected* definition of the graph. To express them, is useful to consider $p = p_n$ dependent on the graph size. Two classical and useful examples of sharp phase transitions in binomial graphs are the ones regarding the connectivity regimes. Here will be presented only the transitions actually used in the thesis. According to the purpose of this work, the propositions stated below will not be proved. Their proofs can be found in [17, 31, 21].

Theorem 2.2. *Let $\mathcal{P}_1$ be the property "there are isolated vertices in $G(n,p_n)$". It has a sharp phase transition w.r.t. the sequence $\frac{\log n}{n}$ of the form*

$$\lim_{n\to+\infty} \mathbb{P}(\mathcal{P}_1 \text{ holds for } G(n,p_n)) = \begin{cases} 0 \text{ if } p_n \geq c\ \frac{\log n}{n},\ c>1 \\ 1 \text{ if } p_n \leq c\ \frac{\log n}{n},\ c<1 \ . \end{cases}$$

Furthermore, the property $\mathcal{P}_1'$: "the graph is connected" has probability of holding on $G(n,p_n)$ that goes to one when n goes to infinity and when

$$p_n \geq c\ \frac{\log n}{n} \quad \text{for } c>1 \ .$$

Theorem 2.3. *Let $\mathcal{P}_2$ be the property "all the connected components in $G(n, p_n)$ have size $O(\log n)$". It has a sharp phase transition w.r.t. the sequence $\frac{1}{n}$ of the form*

$$\lim_{n\to+\infty} \mathbb{P}(\mathcal{P}_2 \text{ holds for } G(n,p_n)) = \begin{cases} 0 \text{ if } p_n \geq \frac{c}{n},\ c>1 \\ 1 \text{ if } p_n \leq \frac{c}{n},\ c<1\ . \end{cases}$$

Remark 2.4. In the case $p_n \geq \frac{c}{n}$ for $c > 1$ it is also proved that with high probability a **giant component** arises, while all other components have size $O(\log n)$.

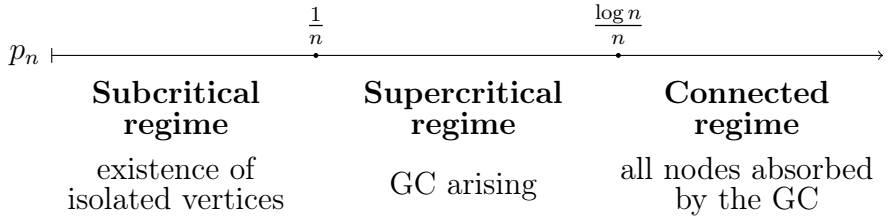


Figure 2.2: ERG connectivity regimes schema.

2.3 Random Geometric Graphs

Now we will introduce the second model of interest, belonging to the family of the *spatial networks*: the Random Geometric Graph (RGG). In general, a spatial network (sometimes also geometric graph) is a graph in which the vertices or edges are spatial elements associated with geometric objects, i.e. the nodes are located in a space equipped with a certain metric [7]. The simplest mathematical realization is a lattice or the RGG itself.

In the following we briefly introduce the basic concepts and results on RGG from [39, 40].

Definition 2.1. A **Random Geometric Graph** $G(n, r)$ is a graph $G = (V, E)$ with $n = |V|$ and vertices distributed in $[0, 1]^d$, $d \geq 1$ independently and uniformly at random, such that a connection between any two pairs of vertices $i = (i_1, \ldots, i_d)$ and $j = (j_1, \ldots, j_d)$ is present with probability one if the Minkowski distance between i and j is lower or equal than a given positive cut-off constant (*radius*) r, i.e. vertices i and j are connected if and only if

$$\left(\sum_{k=1}^{d} |i_k - j_k|^p \right)^{\frac{1}{p}} \leq r\ .$$

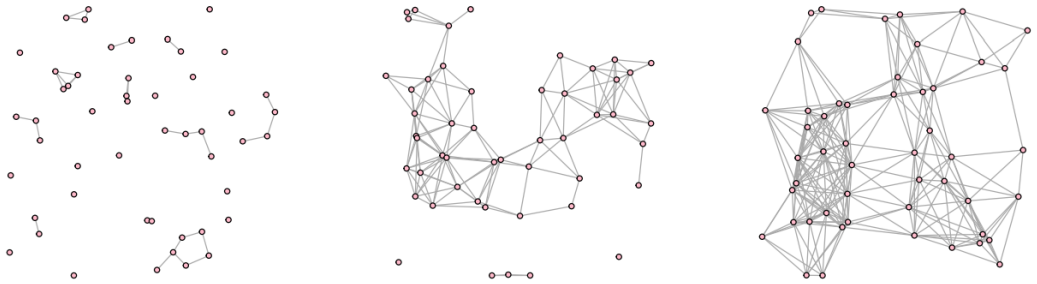


Figure 2.3: Examples of $G(n, r)$ graphs with $n = 50,\ d = 2,\ p = 2$. Starting from the left, the radius assumes the values 0.1, 0.2 and 0.3 respectively.

As clear from the definition just stated, the RGGs are only meant to be defined as *undirected graphs*.

In this work the only RGG model that will be taken into account is the *planar* one, with $d = 2$, and where the used distance is the Euclidean distance, i.e. $p = 2$. This kind of networks were introduced and studied for the first time in [25].

Connectivity regimes. Also the RGGs can be proved to have different stages of connectivity depending on the radius parameter [13, 40]. Let now be $r = r_n$ dependent on the network size.

Proposition 2.2. *Let X be the random variable representing the number of isolated vertices in $G(n, r_n)$. Consider the indicator $X_i = 1$ if i: isolated node, $X_i = 0$ otherwise. Then*

$$\mathbb{E}(X) = \sum_{i:\ node} \mathbb{E}(X_i) = n(1 - \pi r_n^2)^{n-1} \ .$$

Remark 2.3. The mean value computed above can be approximated as

$$\mathbb{E}(X) \sim ne^{-\pi r_n^2 n - O(r_n^4 n)} = \mu e^{-O(r_n^4 n)}$$

where $\mu = ne^{-\pi r_n^2 n}$. How the μ sequence asymptotically behaves characterizes the $G(n, r_n)$ connectivity.

Theorem 2.4. *Let μ be the quantity defined in Remark 2.3. The connectivity of $G(n, r_n)$ is characterized in the following way:*

- *if $\mu \to 0$ then a.a.s.*[1] *the $G(n, r_n)$ is connected;*

[1] a.a.s.: asymptotically almost surely.

- *if* $\mu = \theta(1)$ *then a.a.s. the* $G(n, r_n)$ *the giant component arises;*
- *if* $\mu \to +\infty$ *then a.a.s the* $G(n, r_n)$ *is disconnected.*

Corollary 2.5. *Let* $\mathcal{P}_1$ *be the property "there are isolated vertices in* $G(n, r_n)$*". It has a sharp phase transition w.r.t. the sequence* $\sqrt{\frac{\log n}{\pi n}}$ *of the form*

$$\lim_{n\to+\infty} \mathbb{P}(\mathcal{P}_1 \text{ holds for } G(n, r_n)) = \begin{cases} 0 \text{ if } r_n \geq c\sqrt{\frac{\log n}{\pi n}}, & c > 1 \\ 1 \text{ if } r_n \leq c\sqrt{\frac{\log n}{\pi n}}, & c < 1. \end{cases}$$

Theorem 2.6. *Let* $\mathcal{P}_2$ *be the property "all the connected components in* $G(n, r_n)$ *have size* $O(\log n)$*". It has a sharp phase transition w.r.t. the sequence* $\sqrt{\frac{\lambda_c}{n}}$ *(experimentally* $\lambda_c \sim 2.0736$*) of the form*

$$\lim_{n\to+\infty} \mathbb{P}(\mathcal{P}_2 \text{ holds for } G(n, r_n)) = \begin{cases} 0 \text{ if } r_n \geq c\sqrt{\frac{\lambda_c}{n}}, & c > 1 \\ 1 \text{ if } r_n \leq c\sqrt{\frac{\lambda_c}{n}}, & c < 1. \end{cases}$$

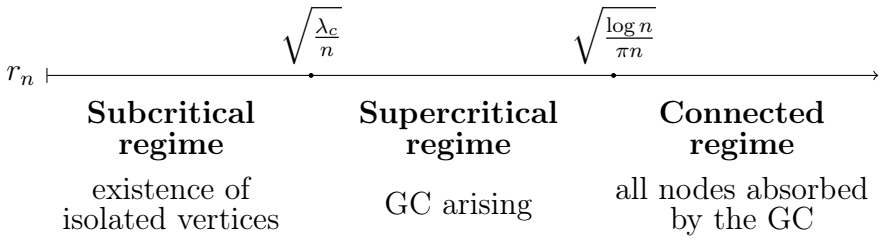


Figure 2.4: RGG connectivity regimes schema.

Chapter 3

LTR on Random Social Graphs

In order to apply the centrality measures introduced in Section 1.3 to random graphs we need to consider them as random social networks, i.e. random networks on which an influence process can be run.

Definition 3.1. A **random social graph** is an influence graph (G, w, f) as defined in Def. 1.1 where G is a random graph, i.e. a graph whose conformation is defined through a probability distribution.

For the graph models introduced in Section 2 the weight function will be constantly equal to one, so from now on we will refer at it as a pair (G, f) since these are the only not trivial elements.

The labeling function f that will be used in this work will always be a *deterministic function*. In particular, only two kind of labeling functions will be used in this thesis.

Definition 3.2. Let G be a random graph on $|V| = n$ vertices. We define the following labeling functions for $i \in V$:

$$f_1(i) = t \cdot (n-1) \; ;$$
$$f_2(i) = t \cdot |pred(i)| \; .$$

where $pred(i)$ will reduce to the $neigh(i)$ set when the graph is undirected and $t \in [0, 1]$.

The f_2 labeling function is coherent with the concept of resistance described in Section 1.2. The f_1 is not the classical definition on threshold used in the LTM, it is an approximation: the number of nodes in the neighbourhood of a specific $i \in V$ is estimated from above by $n - 1$, the maximum number of connections a node can have.

Given the random nature of the graph structures, the quantities computed by Algorithm 1 will be random variables. These will be the quantities measured during the experiment simulations.

All the random variables that will be introduced later on in this section will be defined as functions from the set of vertices, while in fact the set of events they are associated with is the family of possible conformations of the graph that involves a fixed node. This is just an abuse of notation used in order to let the formulation be more understandable.

In the following we write $[n]$ for the set of natural numbers from 0 to n.

Definition 3.3. Let (G, f) be a random social graph on $|V| = n$ vertices. The discrete random variable $level(i, j)$ is defined $\forall i, j \in V$ as

$$level(i,j) : \text{"Level at which } j \text{ is reached starting from } \{i\} \cup neigh(i)\text{"}$$
$$level : V \times V \to [n-1] \ \cup \ \{-1\}.$$

If a node is never reached during the influence expansion the variable assumes the value -1. If a node is in $\{i\} \cup neigh(i)$, it assumes the value zero.

From this definition, two other random variables can be introduced.

Definition 3.4. Let (G, f) be a random social graph on $|V| = n$ vertices. The discrete random variable $maxlevel(i)$ and the real random variable $avglevel(i)$ are defined as:

$$maxlevel(i) = \max_{j \in V} level(i,j), \ \ maxlevel : V \to [n-1] \ ;$$
$$avglevel(i) = \frac{1}{n} \sum_{j=1}^{n} level(i,j), \ \ avglevel : V \to (-1, n-1).$$

Remark 3.5. The $level(i, j)$ random variable corresponds to the level defined in Def. 1.6; the $maxlevel(i)$ corresponds to the $max\_level_{i \cup neigh(i)}$. The average level has no directed correspondence with the quantities defined in Section 1.2, it is the mean of all the levels attribute computed by Alg. 1.

Definition 3.6. Let (G, f) be a random social graph on $|V| = n$ vertices. The discrete random variable $metric(i)$ is defined as:

$$metric(i) : \text{"Number of nodes influenced starting from } \{i\} \cup neigh(i)\text{"}$$
$$metric : V \to [n-1] \setminus \{0\}.$$

Chapter 4

Experiments

The aim of this work is to detect and, if possible, formally prove some property that the LTR has on random social graphs. In particular, the interest is to understand how the definition of the labeling function influences the ranking and check if some of the observed properties shows a phase transition.

Experimental tests on ERG and RGG have been carried with this purpose, in order to explore the behaviour of our metric on different connectivity regimes of the graphs.

4.1 Implementation

All simulation codes are in Python 3. In order to decrease the running time of the simulation, the Numba library[1] and a parallelized implementation[2] have been employed. Since Numba is able to work only with Python Standard Library and with the NumPy[3] library, the adjacency matrix representation has been chosen for the graphs. Some tests have been performed in a preliminary stage to prove that this implementation is more convenient than using adjacency list, SciPy[4] sparse matrix or NetworkX[5] graph structure if the aim is minimize the time consumption. This choice is clearly a trade-off since the used representation is more memory consuming.

Due to the high dimension of the generated data, all the simulation were run on a cluster of computers from the Computer Science Department at UPC[6].

[1] *Numba: A High Performance Python Compiler*, `http://numba.pydata.org/`

[2] Multiprocessing from Python Standard Library, `https://docs.python.org/3/library/multiprocessing.html`

[3] Fundamental package for scientific computing with Python, `https://numpy.org/`

[4] `https://docs.scipy.org/doc/scipy/reference/sparse.html`

[5] *Network Analysis in Python*, `https://networkx.github.io/`

[6] /rdlab, `https://rdlab.cs.upc.edu/`

	Subcritical	Supercritical	Connected
ERG	$\frac{1}{10n}$, $\frac{1}{2n}$	$\frac{2}{3n} + \frac{\log n}{3n}$,	$8 \cdot 10^{-1}$, $7 \cdot 10^{-1}$, $6 \cdot 10^{-1}$,
		$\frac{1}{3n} + \frac{2\log n}{3n}$	$1 \cdot 10^{-1}$, $1 \cdot 10^{-2}$, $x \cdot 10^{-3}$,
			($x = 9$ for $n = 10^3$, $x = 5$ oth.)
RGG	$\sqrt{\frac{1}{10n}}$, $\sqrt{\frac{1}{2n}}$	$\frac{2}{3}\sqrt{\frac{\lambda_c}{n}} + \frac{1}{3}\sqrt{\frac{\log n}{n}}$,	$8 \cdot 10^{-1}$, $7 \cdot 10^{-1}$, $6 \cdot 10^{-1}$,
		$\frac{1}{3}\sqrt{\frac{\lambda_c}{n}} + \frac{2}{3}\sqrt{\frac{\log n}{n}}$	$3 \cdot 10^{-1}$, $1 \cdot 10^{-1}$, $x \cdot 10^{-3}$,
			($x = 6$ for $n = 10^3$, $x = 4$ oth.)

Table 4.1: Random graph parameters per connectivity regimes.

4.1.1 Graphs generation

A set of $G(n, p_n)$ and $G(n, r_n)$ graphs has been considered in order to study the LTR behaviour. Different experiments have been carried on, each of them focusing on a specific aim by changing the parameters of interest. In all of them the different random graphs have been varied by changing the n and the peculiar parameter p_n or r_n (values shown in Tab. 4.1). For each $G(n, p_n)$ graph, both the undirected and directed cases have been analysed (for the directed case, $neigh(i) = succ(i)$); for the $G(n, r_n)$ only the undirected one make sense to be defined. In this work only the images on undirected ERGs will be included because of the high similarity the results on the directed case have shown to them.

Given the phase transitions described in Section 2, the initial probabilities/radii (Table 4.1) have been chosen depending of the current size of the graph. In particular, the probabilities in the *subcritical* and *supercritical regimes* were defined as a formula including the n parameter in order to optimally cover these phases. For the *connected regime*, which represents the largest part of the $(0, 1)$ real interval, the probabilities have been set manually. At an initial stage the value of n only varies in $\{1000, 5000, 10000\}$.

We considered the labeling functions defined in Def. 3.2:

- $f_1(i) = t \cdot (n-1),\ t \in [0, 1]$: this labeling function considers as threshold

n	p_n	Sample size f_1	Sample size f_2
10^3	$p_n < 0.1$	no sample	no sample
	$p_n \geq 0.1$	10^2	10^2
$5 \cdot 10^3$	$p_n < 0.01$	10^3	10^3
	$p_n \geq 0.01$	10^2	10^2
10^4	$p_n < 0.01$	10^3	10^2
	$p_n \geq 0.01$	10^2	10^2

Table 4.2: ERGs sample sizes per probability ranges.

n	r_n	Sample size f_1	Sample size f_2
10^3	$r_n < 0.3$	no sample	no sample
	$r_n \geq 0.3$	no sample	10^2
$5 \cdot 10^3$	$r_n < 0.3$	10^3	10^3
	$r_n \geq 0.3$	10^2	10^2
10^4	$r_n < 0.3$	10^3	10^2
	$r_n \geq 0.3$	10^2	10^2

Table 4.3: RGGs sample sizes per radius ranges.

the percentage t of the theoretical maximum number of neighbours a node can have, i.e. $n-1$. It will be referred as the **max neighbour threshold**. It does not take into account the topology of the graph and it is fixed for every node;

- $f_2(i) = t \cdot |neigh(i)|,\ t \in [0,1]$: this labeling function considers as threshold the percentage t of the number of neighbours a node have. It will be referred as the **neighbour threshold**. It changes for every node and depends on the topology of the graph.

In the fist stages of the experiments, the exploratory ones, the discrete set in which t has been taken is $\{0.25, 0.5, 0.75, 1\}$.

For each influence random graph (G, f) considered in the experiment, we set the number of extracted sample of the graph's probability distribution to $k = 50$ (from now on, referred realizations), in order to better generalize the behaviors observed.

4.1.2 Ranking computation

The simulation of influence expansion outputs a dataset containing for each of the k realizations of the graphs and for each node i:

- `resistance` attribute: the value of t used to obtain the data in the row of the dataset;
- `metric` attribute: the LTR value, not normalized with the graph size. It is the number of influence nodes. Realization of the $metric(i)$ random variable defined in Section 3;
- `max_level` attribute: the maximum expansion level reached during the simulation. Realization of the $maxlevel(i)$ random variable defined in Section 3;
- `avg_level` attribute: the average expansion level of the simulation. Realization of the $avglevel(i)$ random variable defined in Section 3.

The resistance t that defines the value assumed by the labeling function is put as an attribute because of how the influence expansion was implemented. Generated a random graph G, the t parameter is varied at each simulation in order to have different influence graphs. Of course it is not an output but a parameter that defines the input, but it is left in the list just to make the data more readable.

These data are the processed in order to obtain two kind of information: one on each node, computing the average `metric`, `max_level`, `avg_level` values on

all the realizations; the other on the whole influence graph (G, f), calculating the averages of the same attribute on all the realizations and on all the nodes.

In order to ease the LTR computation not all the nodes of the graphs will be used. A sample is extracted by picking a number of nodes equal to a *sample size* variable, by using a uniform distribution on V. The values assigned to this variable change depending labeling function selected. A resume of the sizes used in the numerical simulations are shown in Tab. 4.2 for the ERGs, in Tab. 4.3 for the RGGs.

4.2 First Experiment: max neighbour threshold on ERG

In this Section we will go through the results obtained by the first experiment: the application of the LTR on influence graphs (G, f_1), G: binomial graph. In Phase 1 a preliminary analysis of the results is provided, the parameters used are the basic one described in the Implementation section; on the basis of the observations done at this stage, Phase 2 and Phase 3 will give and discuss the results of the experiments done choosing suitable refinements on the parameters.

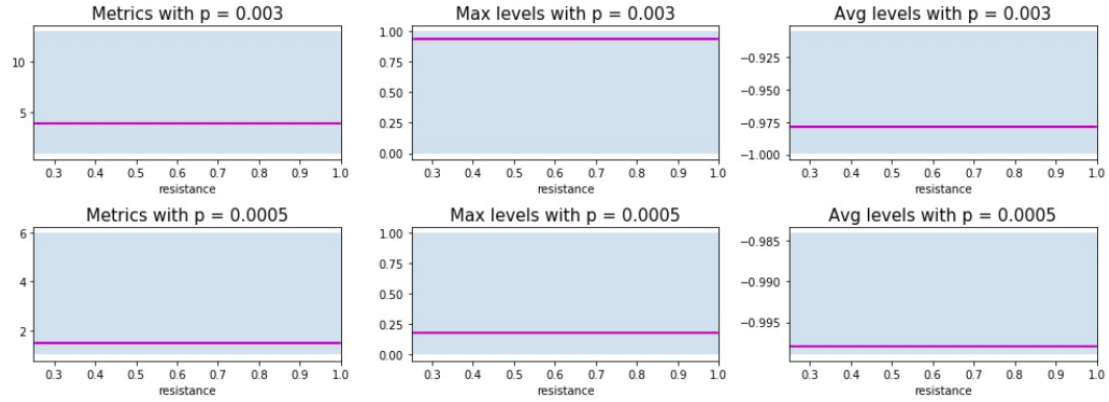
4.2.1 Phase 1

This phase of the first experiment is aimed to get a general idea of how the influence expansion behaves on ERG's different connectivity regimes with f_1 labeling function. Here, only the basic parameters defined in the previous section are used.

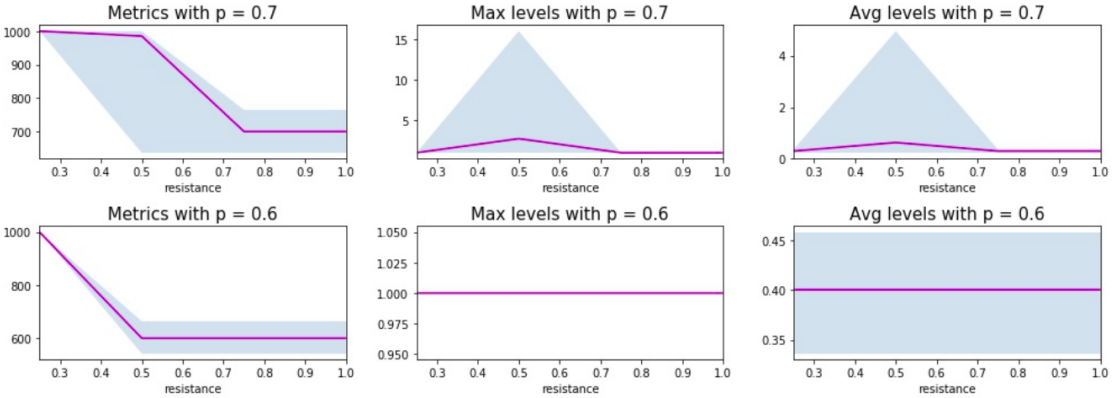
The analysis of the generated data (Appendix A.1) can be resumed in the following observations:

1. `max_level` ≤ 1 almost on all nodes, resistances and probabilities, with some exceptions for high probability values (Figures 4.1a, 4.1b). This indicates that in most of the cases the influence algorithm is not able to activate nodes with a distance from the initial activation set X higher than 1 (coherent with the results obtained in [23] about the BOLTR). For $p_n < 0.6$ it happens because the probability value is too low to have enough links; in the other cases, because the graph is highly connected and the nodes are reached in time 1 or never reached;

2. `avg_level` ≤ 0 for probabilities before the connected regime or above but near the critical point and for all nodes (Figure 4.1a), which indicates that the number of inactive and never reached nodes ($level = -1$) is high with respect to the number of activated ones;



(a) Example for the Observations 1 to 4.



(b) Example for the Observations 1, 4 and 5.

Figure 4.1: Undirected case, $n = 1000$. On the x axis the values of t used (not the complete interval, just Phase 1 discrete set); on the y axis the values assumed by the parameter in the title.

3. `max_level` $= 0$ for at least one node for probabilities before the connected regime (Figure 4.1a), which means that there are no activated nodes outside of the initial set $\{i\} \cup succ(i)$;

4. the parameter t seems to have a low influence on how the algorithm behaves (Figures 4.1a, 4.1b). This is noticeable by the fact that almost all the results show a mean value represented as a straight line. This is explicable by the hard requirement f_1 represents, being defined as a percentage of maximum number of node an actor can be connected with. A difference is only visible for probability values close to 1, which means that the expected degree $\mathbb{E}(d(i)) = (n-1)p_n$ (Eq. (2.1)) of each node is near to the value $n-1$, so it is expected to have more nodes with the structural possibility to be ac-

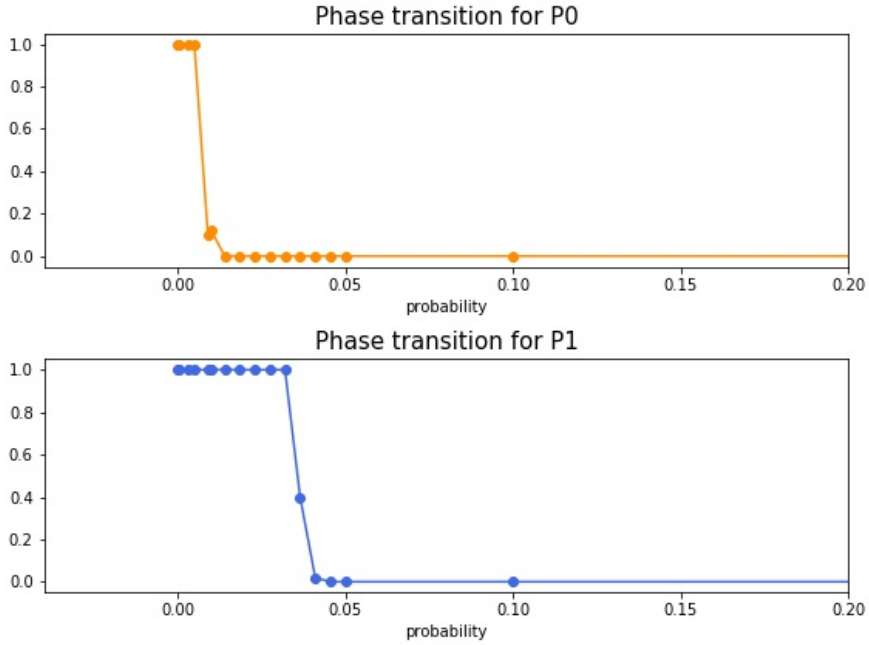


Figure 4.2: Phase transitions for Properties 0 and 1, average truth values on the y axis. Data shown for $t = 0.5$, $n = 10^3$, max neighbours undirected case.

tivated. Another explanation, which does not exclude the previous one but complete it, is that when $d(i) \sim \mathbb{E}(d(i))$ the value of t impedes the expansion to proceed outside the initial activation set when $t > p_n$;

5. `metric` shows a change of behaviour for high probability values and $t \in [0.25, 0.5]$ (Figure 4.1b), coherent with the inflection point observed in [16]. In correspondence of the changing point, the `max_level` variable assumes values ≥ 1 for $p_n = 0.7$.

These first patterns allow the next simulation to be organized in the following way: Phase 2 will be aimed to define and deeper explore the observation about the levels (`max_level` and `avg_level`); Phase 3 will analyse the behaviour of the `metric` parameter.

4.2.2 Phase 2

Looking at the results given by the Phase 1, two properties about the `max_level` and `avg_level` parameters can be formulated.

In order to analyse the behaviour of this property and see if it possible to detect a phase transition, some new graphs have been generated with $p_n \in [0.01, 0.05]$ and n varied in $[10^3, 10^4]$ with steps of 10^3.

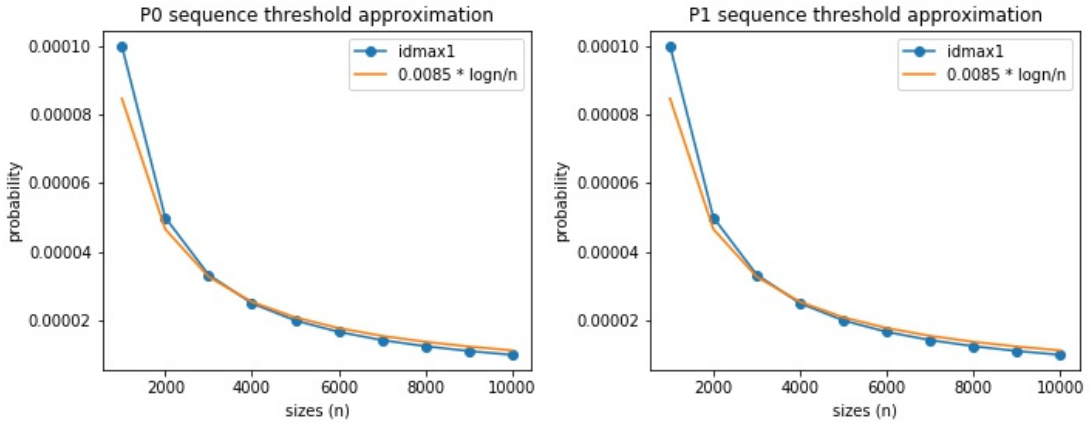


Figure 4.3: Threshold sequences possible formulas. Data shown for $t = 0.5$, represented the last probability values for which the property holds in every realization.

Property 0. The Observation 3 can be formalized as

$$\mathcal{P}_0 : \text{"}\exists i \mid \texttt{max\_level}(i) = 0\text{"}$$

which means that, after the simulation of influence expansion started from node i, the only actors resulting activated are the ones in the initial set.

In Appendix A.2.1 and A.2.3 the images representing the results are shown: in the former set of images the truth values assumed by the property are displayed versus the corresponding probability values; in the latter, it can be seen an approximation of the threshold sequence of the phase transition. A representative example is reported in Figure 4.2. It is clear from these representation of the data that the t parameter is not influential at all.

In Figure 4.3 a possible formulation of the threshold sequence is compared with the data. Looking at the picture on the left, the one representing the last probability values for which the property holds in every realization for $\mathcal{P}_0$, a clear shape of $\Theta(\frac{\log n}{n})$ can be noticed. This function will be the threshold guess we will prove in the next section.

Property 1. The Observations 1 and 2 can be combined in a unique property

$$\mathcal{P}_1 : \text{"}\exists i \mid \texttt{max\_level}(i) \leq 1 \ \wedge \ \texttt{avg\_level}(i) \leq 0\text{"}$$

where $i \in V$: node of the graph. In this way we are capturing the inability of the influence to be spread along the graph when the probability value is low.

The results about the phase transition existence and its possible threshold sequence are represented in Appendix A.2.2 and A.2.3. It is clearly shown a sharp phase transition, that seems to start being false for bigger values of p_n. This

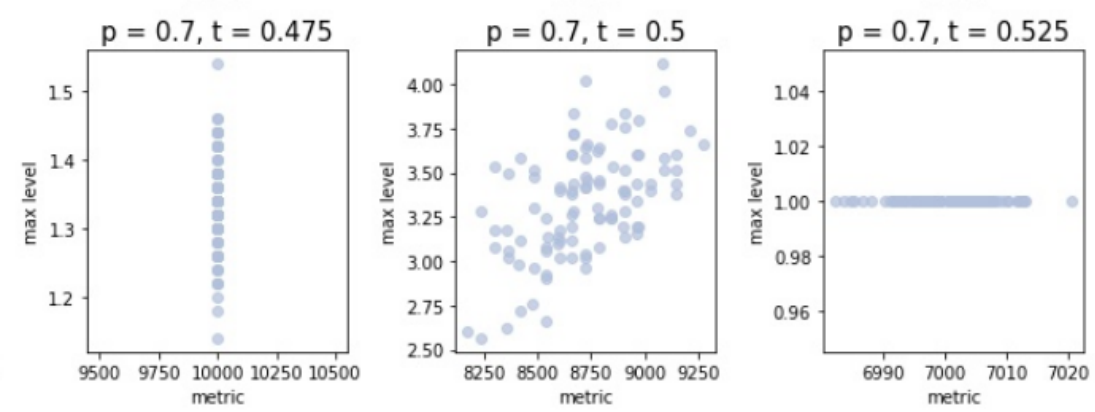


Figure 4.4: Behaviour of the metric and the maximum level parameters around the inflection point, shown for $n = 10000$.

indicates that $\mathcal{P}_1$ probably has a wider transition gap with respect the property $\mathcal{P}_0$. In addition, it can be seen that even in this case the transition seems to be independent from the value of t.

In Figure 4.3 the data-estimated transition threshold is compared with some functions. The image on the right shows an identical behaviour to the one of the $\mathcal{P}_0$ property, so the threshold guess will be the same logarithmic sequence $\Theta(\frac{\log n}{n})$.

4.2.3 Phase 3

The Observation 5 of the Phase 1 does not give enough information to formulate any kind of hypothesis about the inflection point behaviour and the corresponding peak in `max_level` detected for $p_n = 0.7$. It can be better explored: it is probable that, for probability values which are high enough to activate the whole graph, there exists a value t for which an inflection point is observed and `max_level` ≥ 1.

A refinement of the t values has been done by covering the $[0.2, 1]$ interval with discrete values at distance 0.025 for probability values $p_n \in \{0.6, 0.7, 0.8\}$.

The images related to this simulations are visible in Appendix A.1.2: in the first set of images it is clear that for each considered size there exists a inflection point almost independent from n but strongly dependent on the probability parameter p_n, increasing with its values.

It can be noticed in Appendix A.1.2 that the inflection point of the metric is characterized by the different behaviours depending on t (zoom on Figure 4.4):

1. an initial phase in which the metric value is still maximum but assumed after different time steps of the influence expansion. In this phase a vertical line appears in the plots;

2. a critical phase in which the algorithm finds more difficult to spread the influence (the metric value is not always maximum) but the maximum level reached has a peak;

3. a final phase in which the metric value is more stable and the maximum levels decrease again to one. The former oscillates around a high value of influenced nodes different from the maximum, the latter becomes stable at one.

The final phase is reached for $t > p_n$. This Observation will be theoretically explained in the next chapter.

The presence of an inflection point for the metric was already know from [16], even if only on real-world data. What it is interesting to notice here is that for this specific model of random graph and this labeling function, the inflection point is detected in a different range of t: on the data analysed in the cited work, the range was [0.2, 0.5], while in this case for the probabilities 0.6, 0.7, 0.8 we found respectively the ranges [0.3, 0.4], [0.45, 0.55], [0.6, 0.7].

In addition, an unexpected behaviour of the `max_level` parameter has been noticed around these point.

4.2.4 Notes on the directed case

This paragraph is meant to show in brief that the results obtained for the undirected random social graphs are valid also for the directed ones, with the only difference that in the second case we do not have any known connectivity phase transition.

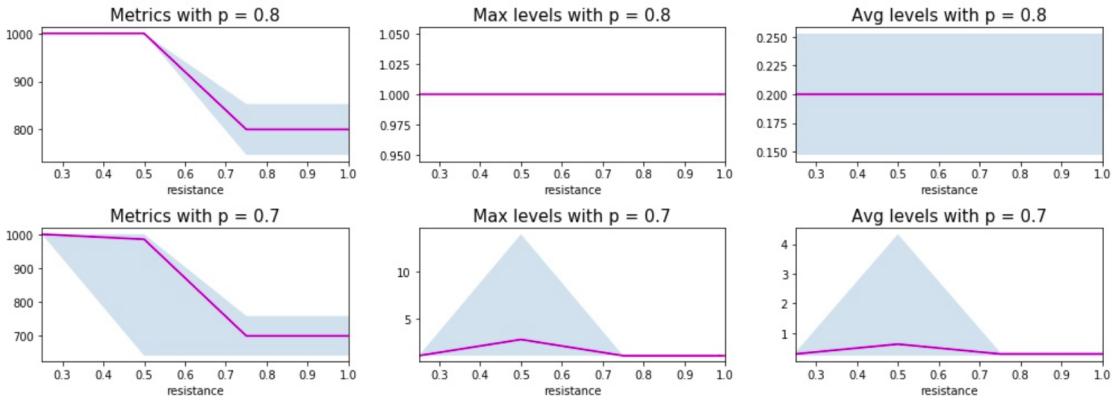
We can clearly see in Figures 4.5b and 4.5a that the observations done in Phase 1 on the undirected (G, f_1) are still valid. It can also be seen that the properties 0 and 1 show the same phase transitions.

About the inflection point and the three phases noticed when considering the relation between `metric` and `max_level`, they do not present any difference. Moreover, the `metric` data seem to have a Gaussian distribution.

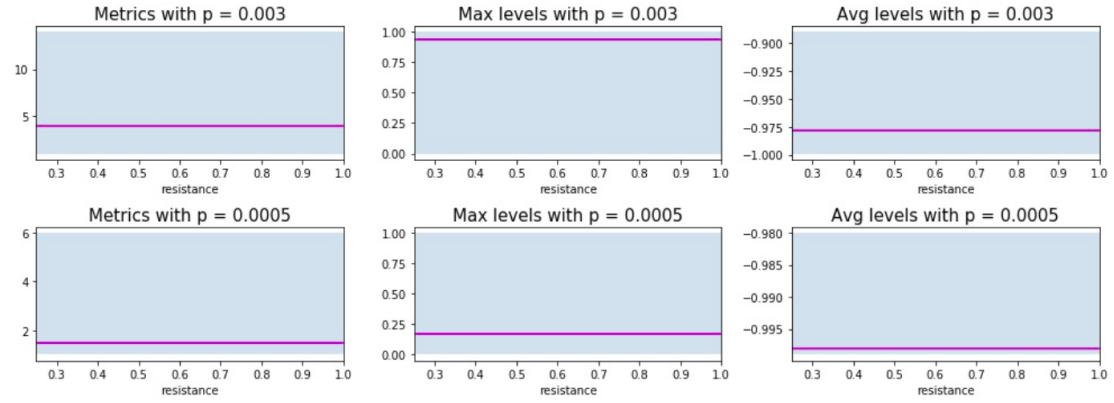
4.3 Second Experiment: neighbour threshold on ERG

In this Section we will go through the results obtained by the second experiment: the application of the LTR on influence graphs (G, f_2), G: binomial graph.

In Phase 1 a preliminary analysis of the results is provided, when the parameters used are the basic ones described in the Implementation section; on the basis of the observations done at this stage, Phase 2 and Phase 3 will discuss the other experiments based on refinements of the parameters.



(a) Low probabilities behaviour.



(b) High probabilities behaviour.

Figure 4.5: Directed case, $n = 1000$. On the x axis the values of t used (not the complete interval, just Phase 1 discrete set); on the y axis the values assumed by the parameter in the title.

4.3.1 Phase 1

In this second framework, the Phase 1 is again exploratory. The influence expansion is performed on ERG's different connectivity regimes with f_2 labeling function. As before, only the basic parameters defined in the previous section are used. Note that for $\mathbb{E}_{i,k}$ we are computing the average in the data, i.e. sample average, not the theoretical one.

The properties arising from data (Appendix A.3) can be resumed as:

1. the `max_level` variable behaviour is not as stable as observed in previous experiment. Here the observation $\mathbb{E}_{i,k}(\texttt{max\_level}) < 1$ is true for t big enough (in general, $t \geq 0.5$) and for all the probability values (Figures 4.6a, 4.6b).

As regards the variability of this parameters, is high in the low probability values and strongly depends on t;

2. `avg_level` ≤ 0 for probabilities before the connected regime, for t big enough and for all nodes (Figure 4.6a). This indicates that the number of inactive and never reached nodes ($level = -1$) is high with respect to the number of activated ones;

3. $\mathbb{E}_{i,k}$(`max_level`) ≈ 0 for probabilities before the connected regime and t big enough (in general, $t \geq 0.5$), which means that there are no activated nodes outside the initial set $\{i\} \cup succ(i)$ (Figure 4.6a);

4. the parameter t here has influence on the quantities observed (Figures 4.6a, 4.6b). Around $t = 0.5$ the $\mathbb{E}_{i,k}$(`max_level`) and $\mathbb{E}_{i,k}$(`avg_level`) variable show an inflection point for $p_n < 0.6$ and outside the subcritical regime. Above $p_n = 0.6$, in the highly connected regime, the behaviour is not clear and should be better explored;

5. `metric` shows a change of behaviour for high probability values and $t \in [0.5, 0.8]$, probably connected with the observation 4 (Figure 4.6b). In addition, especially for $n = 10^3$ it seems that for small values of t the Giant Component is totally (or nearly totally) activated.

These observations are similar but less regular with respect t as the ones noticed in the First experiment.

4.3.2 Phase 2

Starting from the observation done during Phase 1, the same properties defined for the First Experiment can be analysed again here.

Property 0. The Observation 3 can be formalized as

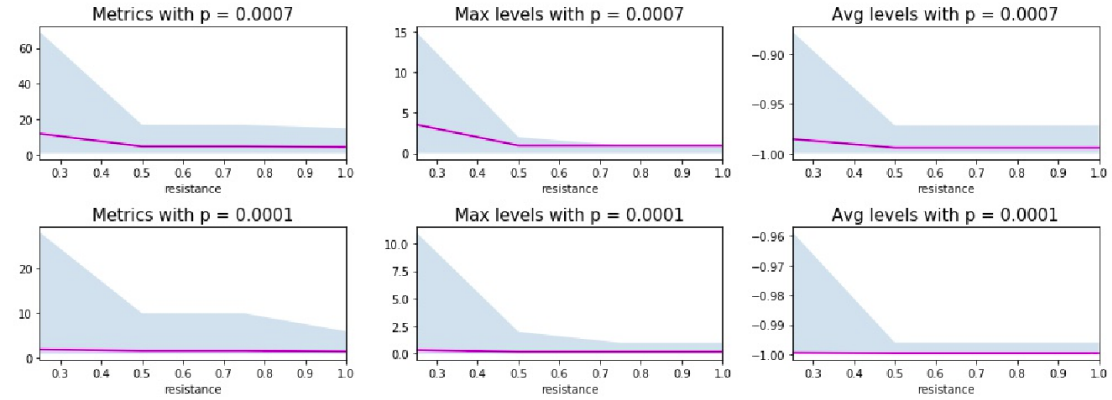
$$\mathcal{P}_0 : \text{"}\exists i \mid \texttt{max\_level}(i) = 0\text{"}.$$

Property 1. The Observations 1 and 2 can be combined in a unique property

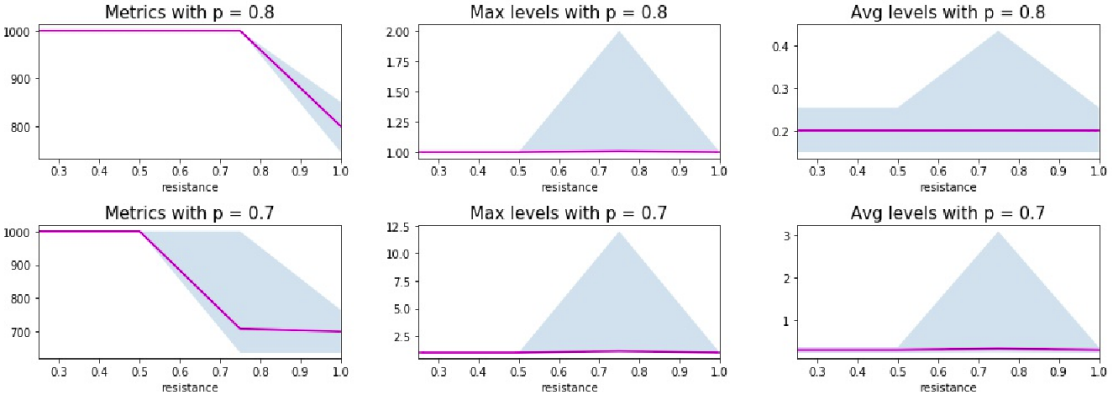
$$\mathcal{P}_1 : \text{"}\exists i \mid \texttt{max\_level}(i) \leq 1 \ \wedge \ \texttt{avg\_level}(i) \leq 0\text{"}.$$

In order to analyse the behaviour of this property and see whether it possible to detect a phase transition, some new graphs have been generated with $p_n \in [0.01, 0.05]$.

The refinement shows for $\mathcal{P}_0$ and $\mathcal{P}_1$ the same phase transitions already observed in the previous experiment (see Appendix A.4.1, focus on 4.7). This was of



(a) Example for the Observations 1 to 4.



(b) Example for the Observations 1, 4 and 5.

Figure 4.6: Undirected case, $n = 5000$ (a), $n = 1000$ (b). On the x axis the values of t used (not the complete interval, just Phase 1 discrete set); on the y axis the values assumed by the parameter in the title.

course expected for $\mathcal{P}_0$ since the step 0 of the influence expansion does not depend on the resistance function defined on the node. It was less obvious for the $\mathcal{P}_1$, that could have been more dependent on the labeling function and on the t value.

Because of this similarity, we will not repeat the discussion about the phase transition sequence threshold, which was already carried out in the Phase 2 of the First Experiment (Section 4.2).

A different property seems to arise in the supercritical phase when the labeling function is f_2, as underlined in Observation 5.

Property 2.

$$\mathcal{P}_2 : \text{"}\exists i \mid \texttt{metric}_i = |GC|\text{"}.$$

This property is not studied in this thesis, but it is left as a possible future

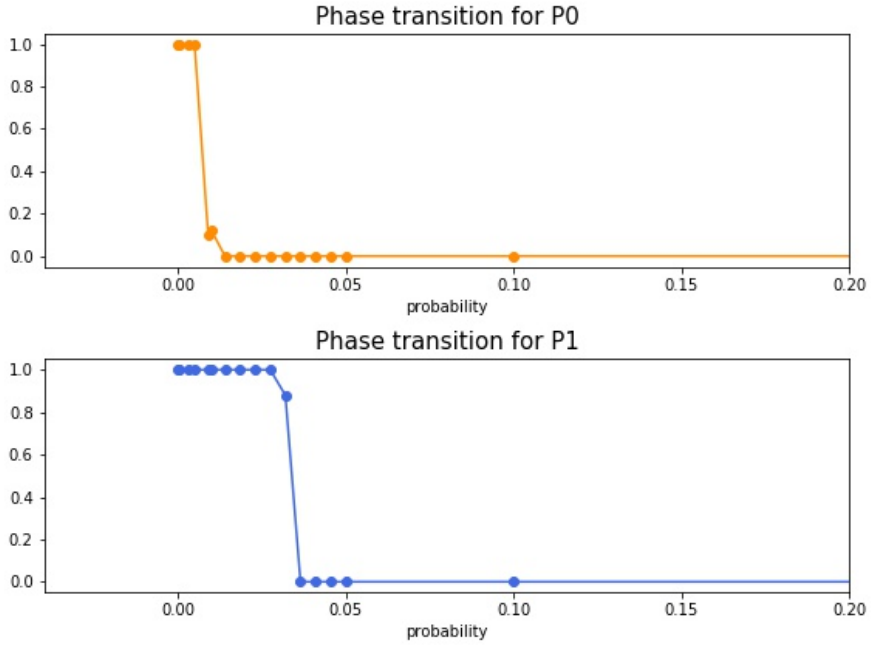


Figure 4.7: Phase transitions for Properties 0 and 1, average truth values. Data shown for $t = 0.5$, $n = 10^3$, max neighbours undirected case.

work.

4.3.3 Phase 3

This phase of refinement is based on the Observation 5 of Phase 1, in particular it is focusing on the inflection point shown only by the simulations on $n = 10^3$. The data generated in the first phase are not informative since the values chosen for the t parameter are not specific enough to identify some behaviour when the LTR metric values decrease.

Picking new values for t of the form 0.x5, where x $\in \{2, 3, 4, 5, 6, 7, 8, 9\}$, the data showed an inflection point for the `metric` with a behaviour close to the one already observed for the previous experiment.

In addition, the same pattern of evolution of the relationship between the `metric` and `max_level` parameters shown in Experiment 1 can be observed here. They can be seen in Appendix A.3.2, while a focus is reported in Figure 4.8. For time reasons, the refinement is here less precise: the three phases already described in detail in Section 4.2.3 (initial, critical, final) are still visible but with a worst resolution. However, we can easily infer that the behaviour may be exactly the same.

The ranges in which the three phases are observed are the following: $[0.75, 0.85]$

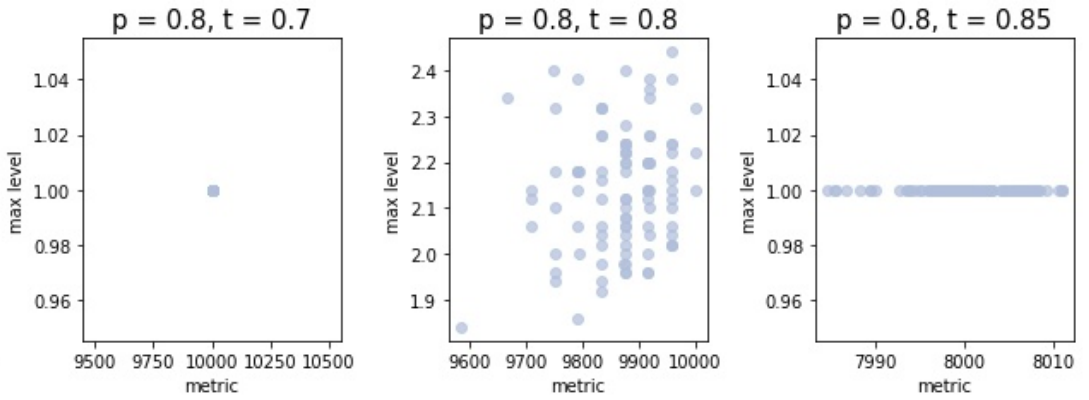


Figure 4.8: Behaviour of the metric and the maximum level parameters around the inflection point, shown for $n = 10000$.

for $p_n = 0.8$, $[0.85, 0.75]$ for $p_n = 0.7$, $[0.75, 0.65]$ for $p_n = 0.6$. They are different from the ones observed in [16] on real-world data and also different from the results obtained from the First Experiment: the ranges have extremes greater then in the previous case, but of a similar size.

Again it can be noticed that the final phase, in which all the nodes are activated at level 1 but the LTR is not maximum, is reached for $t > p_n$.

4.3.4 Notes on the directed case

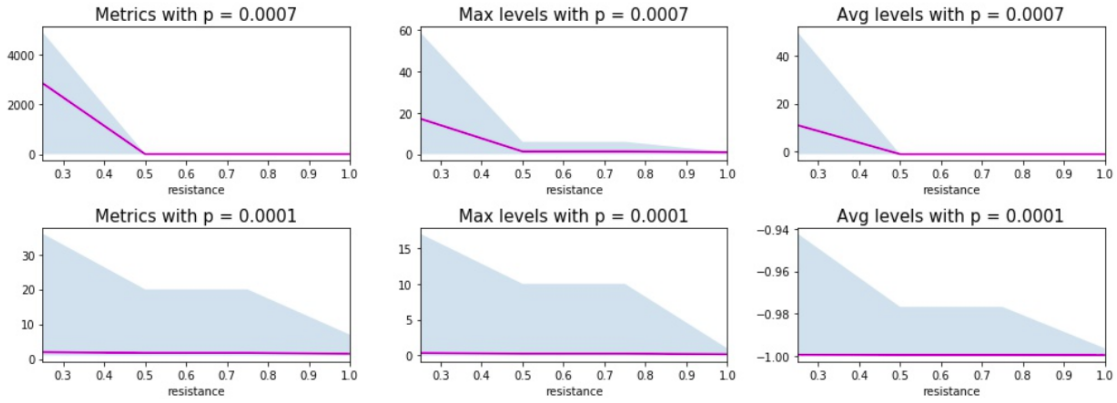
In this paragraph we will comparing the results on undirected random social graphs and the directed ones. We can observe that the results are exactly the same apart from one fact: the properties noticed in the directed graphs cannot be related to connectivity regimes of the network since the results on the phase transitions are valid only in the undirected case.

In Figures 4.9b and 4.9a the same pictures analysed in Phase 1 are reported, but for the directed case. It is clear that the observations already done in Phase 1 are still valid. It can also be seen that the properties 0 and 1 show the same phase transitions.

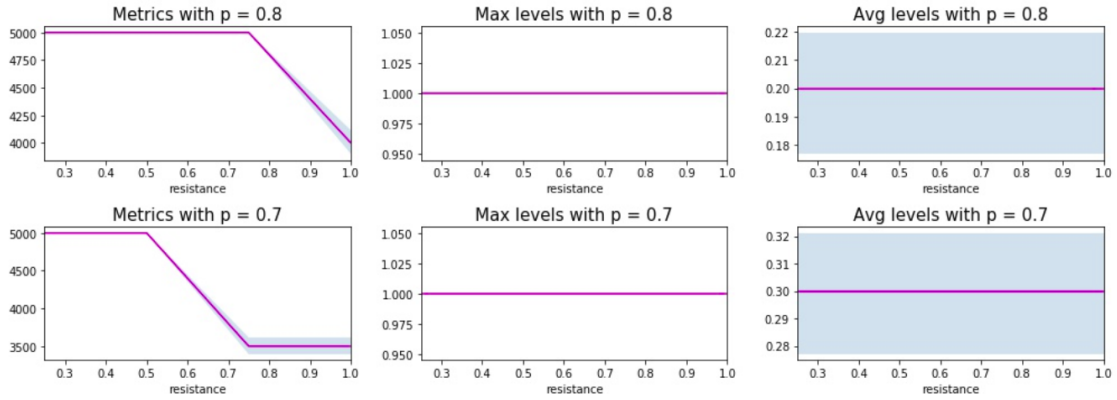
About the inflection point and the three phases noticed when considering the relation between `metric` and `max_level`, they do not present any difference. Moreover, the `metric` data seem to have a Gaussian distribution.

4.4 Third Experiment: max neighbour threshold on RGG

In this Section we will analyse the results obtained by the third experiment: the application of the LTR on influence graphs (G, f_1), G: random geometric graph.



(a) Low probabilities behaviour.



(b) High probabilities behaviour.

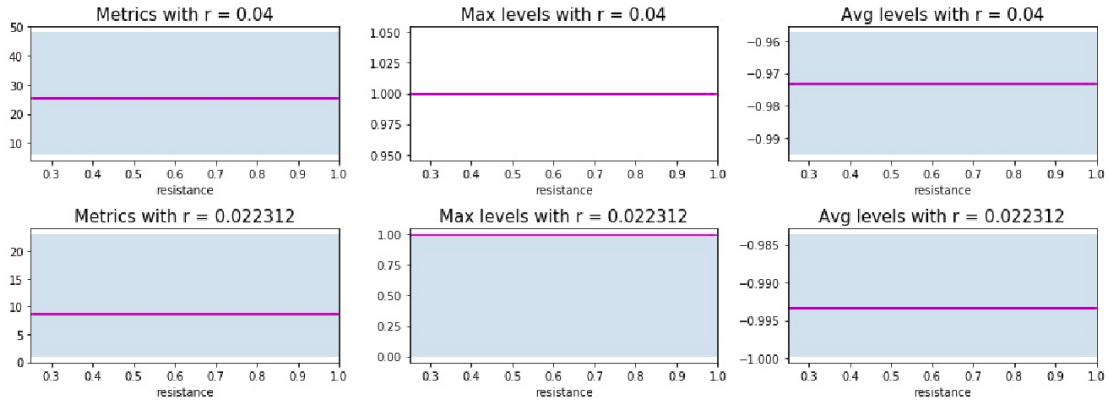
Figure 4.9: Directed case, $n = 1000$. On the x axis the values of t used (not the complete interval, just Phase 1 discrete set); on the y axis the values assumed by the parameter in the title.

The simulations run on this model has just the aim of give a general idea on how the LTR performs on RGGs. Because of this, just the initial parameters described in Section 4.1.2 will be used in all the phases.

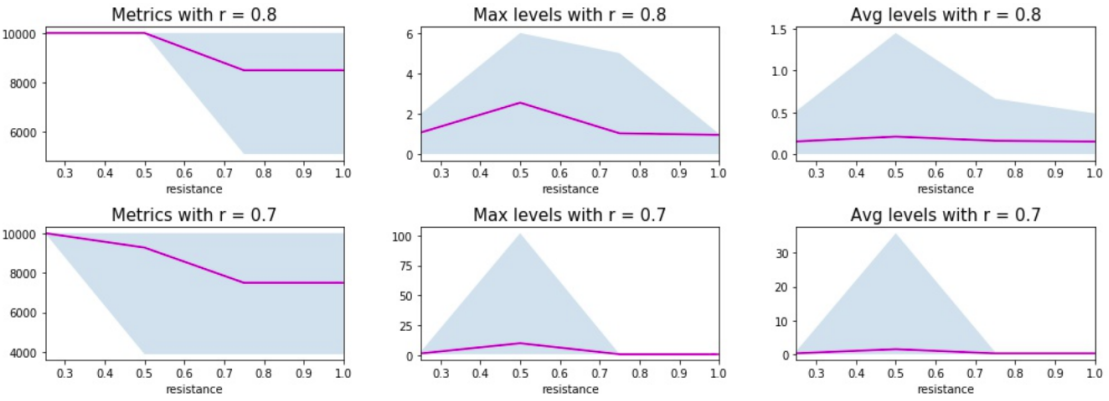
4.4.1 Phase 1

The third experiment's exploratory phase gave the results shown in Appendix B.1. The following list resumes the main features observed:

1. `max_level` ≤ 1 almost on all nodes, before connected regime and for the first radii of the connected phase (Figures 4.10a, 4.10b). Again, the algorithm of expansion shows a difficulty to reach nodes far from the initial set (coherent with the results obtained in [23] about the BOLTR);



(a) Example for the Observations 1 to 4, $n = 5000$.



(b) Example for the Observations 1, 4 and 5, $n = 10000$.

Figure 4.10: Parameter behaviour examples. On the x axis the values of t used (not the complete interval, just Phase 1 discrete set); on the y axis the values assumed by the parameter in the title.

2. `avg_level` ≤ 0 before connected regime and for the first radii of the connected phase, on all nodes (Figure 4.10a). It indicates that the number of inactive and never-reached nodes ($level = -1$) is high with respect to the number of activated ones;

3. `max_level` $= 0$ for at least one node for probabilities before the connected regime (Figure 4.10a), which means that there are no activated nodes outside the initial set $\{i\} \cup neigh(i)$. The mean value of `max_level` is always above zero before high r_n. After, it is zero where it is not shown a peak;

4. the parameter t have a low influence on how the algorithm behaves when $r_n < 0.3$ (Figures 4.10a, 4.10b). Almost all the results show a mean value represented as a straight line. Again we recall that f_1 represents a hard

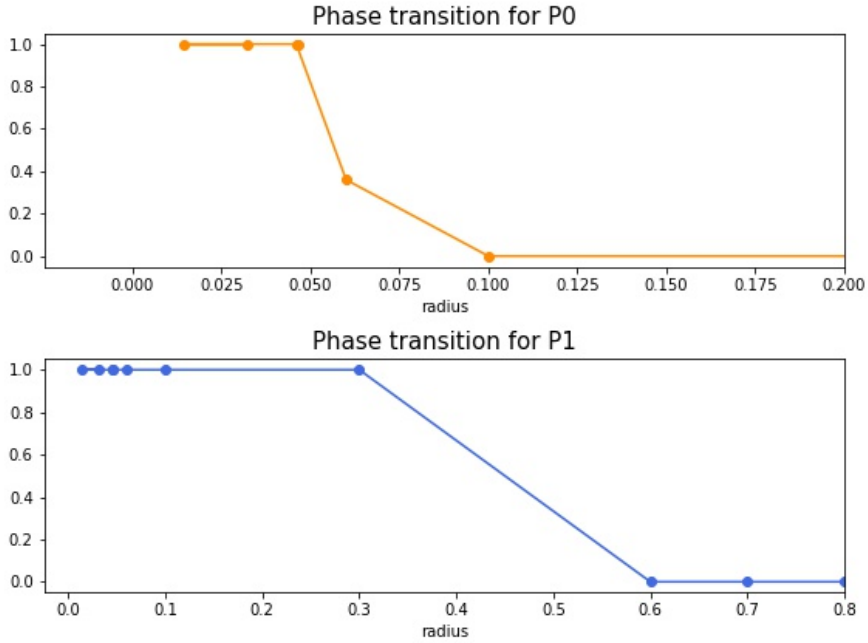


Figure 4.11: Phase transitions for Properties 0 and 1, average truth values on the y axis. Data shown for $t = 0.5$, $n = 10^3$, max neighbours case.

condition, being defined as a percentage of maximum number of node an actor can be connected with;

5. `metric` shows an inflection for high radii values, $r_n > 0.6$ (Figure 4.1b). For $n = 10^3$, this is true also for $r_n = 0.3$ which suggests that the connected regimes could be better explored to find that the inflection point appears for a wider range of radii. In correspondence to the changing point, the `max_level` variable assumes values ≥ 1.

These observations are similar to the ones made for the First experiment, with some slight differences. It can be underlined that even for a different model, f_1 does not behave so differently depending on the value of t.

4.4.2 Phase 2

Since Phase 1 underlined a behaviour similar to the one observed in Experiments 1 and 2, we can try to analyse the same properties.

Property 0. The Observation 3 can be formalized as

$$\mathcal{P}_0 : \text{"}\exists i \mid \texttt{max\_level}(i) = 0\text{"}.$$

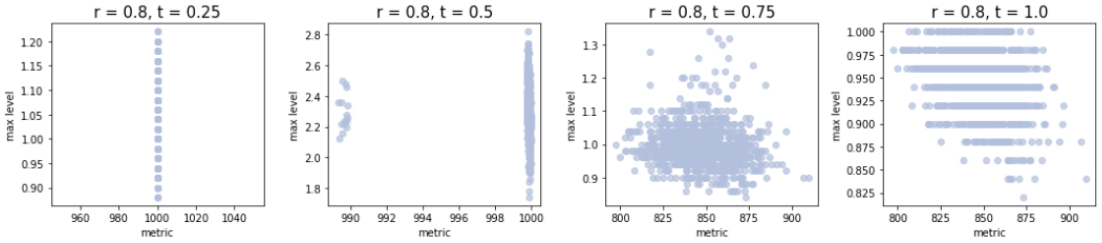


Figure 4.12: Behaviour of the metric and the maximum level parameters around the inflection point, shown for $n = 10000$.

Property 1. The Observations 1 and 2 can be combined in a unique property

$$\mathcal{P}_1 : \text{"}\exists i \mid \texttt{max\_level}(i) \leq 1 \ \wedge \ \texttt{avg\_level}(i) \leq 0\text{"}.$$

Even without the refinements on the radii values, the properties clearly show a phase transition. All the results can be seen in Appendices B.2.1 and B.2.2, while an example in shown in Figure 4.11.

The $\mathcal{P}_0$ property is for sure not true in the connected regime. As we will later see, it is only dependent on the structure of the graph instead of being influenced also by the labeling function. This allows us to conjecture that $\mathcal{P}_0$ on RGGs will have an asymptotic behaviour similar or equal to the one shown on ERGs.

As regards $\mathcal{P}_1$, from the data available now it seems to have a threshold sequence belonging to the connected regime: the radii 0.1 and 0.3 have truth values always equal to True on all the data generated by us. So, it seems that this property is not equivalent or similar to any of the already existing ones on connectivity.

Given the low number of radii considered in these phase, we will not try to graphically guess the values of the threshold sequences for the two properties.

4.4.3 Phase 3

This phase is about the Observation 5 of Phase 1: the inflection point shown by the `metric` parameter for high radii values. The full results can be seen in Appendix B.1.

As done in the previous experiments, it has been looked carefully to the relation between the `metric` and the `max_level` variables. In Figure 4.12 we can see that the pattern initial phase - critical phase - final phase is almost equal to the ones shown in ERGs. It is worth mentioning that, for $r_n = 0.8$ and every n, even in $t = 1$ the completely straight behaviour is not reached. However, the `max_level` is never higher than one and there some kind of organization in horizontal lines.

A refinement on the t parameters may surely reveal the very same pattern observed in the previous experiments, so that some intervals for the inflection point could be observed. It is left as future work.

4.5 Fourth Experiment: neighbour threshold on RGG

In this Section we present and discuss the results of the fourth experiment: the application of the LTR on influence graphs (G, f_2), G: random geometric graph.

The simulations run on this model has just the aim of giving a general idea on how the LTR performs on RGGs. Because of this, just Phase 1 is performed and all the observations done are left to be deeper analysed in future works.

4.5.1 Phase 1

Now we will comment the preliminary results of the Fourth experiment, fully available in Appendix B.3:

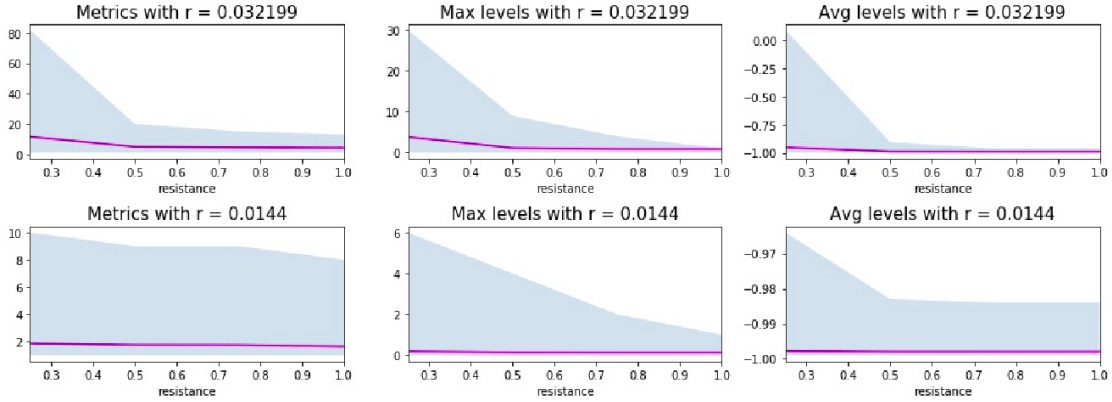
1. the `max_level` is always near zero in mean, apart from the (low) peaks shown for high radii and for small values of the parameter t ($t \leq 0.5$). Examples of this observation can be seen in Figures 4.13a, 4.13b. As regards the variability of this parameters, is high in the low probability values and strongly depends on t;

2. `avg_level` ≤ 0 in the subcritical regime of connectivity, for all t and for all nodes (Figure 4.13a). This indicates that the number of inactive and never-reached nodes ($level = -1$) is high with respect to the number of activated ones. In some cases, almost all the nodes are never visited since the mean `avg_level` is near -1.In the other regimes, we always observe `avg_level` ≥ 0;

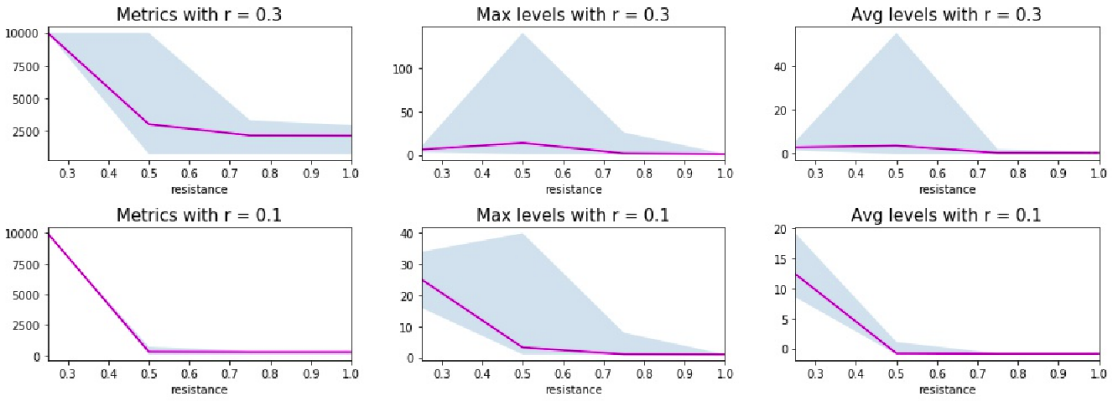
3. the parameter t here has influence on the quantities observed (Figures 4.13a, 4.13b). Around $t = 0.5$ the $\mathbb{E}_{i,k}$(`max_level`) and $\mathbb{E}_{i,k}$(`avg_level`) variable show an inflection point for $r_n \leq 0.3$. Strictly above $r_n = 0.3$, in the highly connected regime, the mean values cited before have low peaks correspondent to various values of t;

4. `metric` clearly shows the presence of an inflection point when the radii is in the connected regime (Figure 4.13b). In addition, it seems that for small values of t the Giant Component is totally (or almost totally) activated.

Note that for $\mathbb{E}_{i,k}$ we are computing the average in the data, i.e. sample average, not the theoretical one.



(a) Example for the Observations 1 to 3, $n = 1000$.



(b) Example for the Observations 1, 3 and 4, $n = 10000$.

Figure 4.13: Parameter behaviour examples. On the x axis the values of t used (not the complete interval, just phase 1 discrete set); on the y axis the values assumed by the parameter in the title.

4.5.2 Phase 2

The data analysed in Phase 1 show a general behaviour in common with all the other experiment, hence we will analyse the same properties.

Property 0.

$$\mathcal{P}_0 : \text{"}\exists i \mid \texttt{max\_level}(i) = 0\text{"}.$$

Property 1.

$$\mathcal{P}_1 : \text{"}\exists i \mid \texttt{max\_level}(i) \leq 1 \ \wedge \ \texttt{avg\_level}(i) \leq 0\text{"}.$$

Property 1 could have been defined differently since $\texttt{max\_level}(i)$ always assumes values near zero in mean (Observation 1), but in order to do a consistent comparison with all the other experiments we decided to not change it.

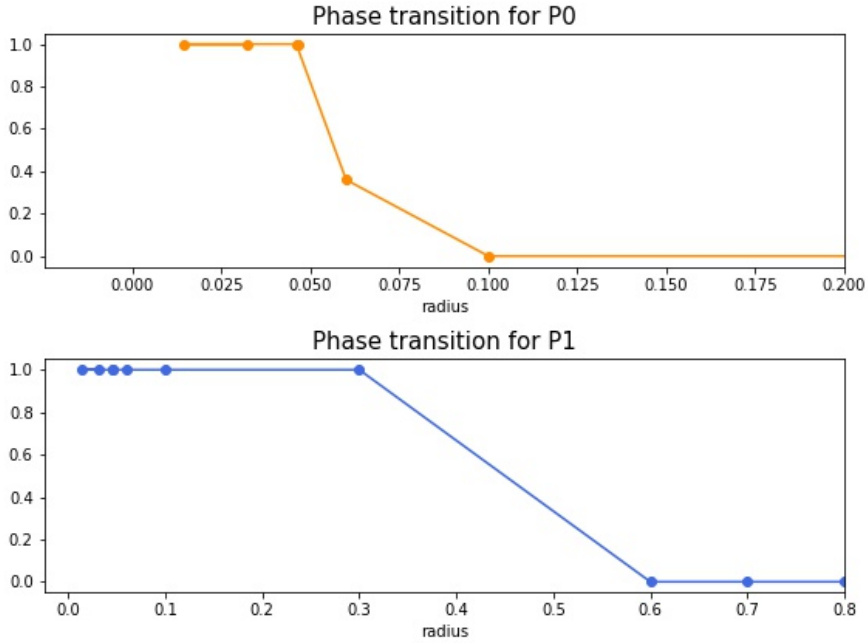


Figure 4.14: Phase transitions for Properties 0 and 1, average truth values on the y axis. Data shown for $t = 0.5$, $n = 10^3$, max neighbours case.

As we can see in Figure 4.14, the $\mathcal{P}_0$ property has the expected behaviour: the phase transition clearly happens in the same way as we saw in Experiment 3. It is again independent from t and assumes value False in the connected regime.

Property $\mathcal{P}_1$ behaves in the same way as in the previous experiment for $t > 0.25$, i.e. it shows a phase transition in the connected regime, but has a peculiarity: for $t = 0.25$ we can observe in Figure 4.15 that the transition has a different shape, the change of the truth value appears near the isolated regime. Then, we can affirm that $\mathcal{P}_1$ somehow depends on t on random social graphs (G, f_2), G : RGG.

Given Observation 4, another Property can be defined as done in the Second Experiment:

Property 2.

$$\mathcal{P}_2 : \text{"}\exists i \mid \texttt{metric}_i = |GC|\text{"}.$$

Again, the analysis of this property is left as a possible future work.

4.5.3 Phase 3

Here the Observation 4 of Phase 1 will be deeper analysed. In particular, an inflection point for the `metric` parameter is noticed for every radius in the

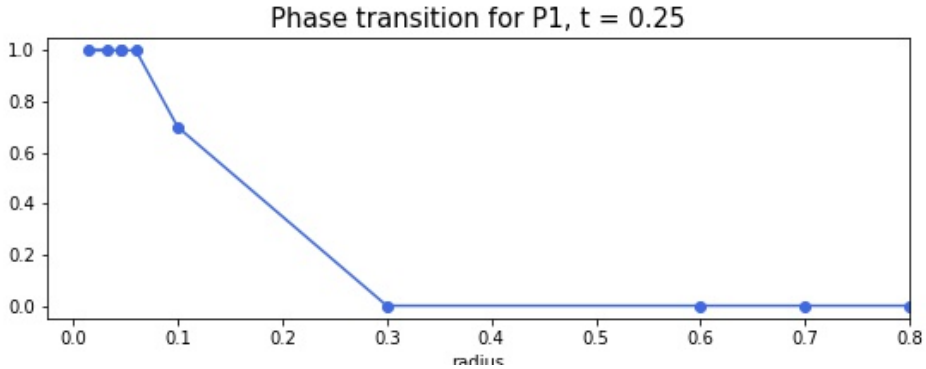


Figure 4.15: Phase transition for Property 1 and $t = 0.25$, average truth values on the y axis. Data shown for $n = 10^3$, neighbours case.

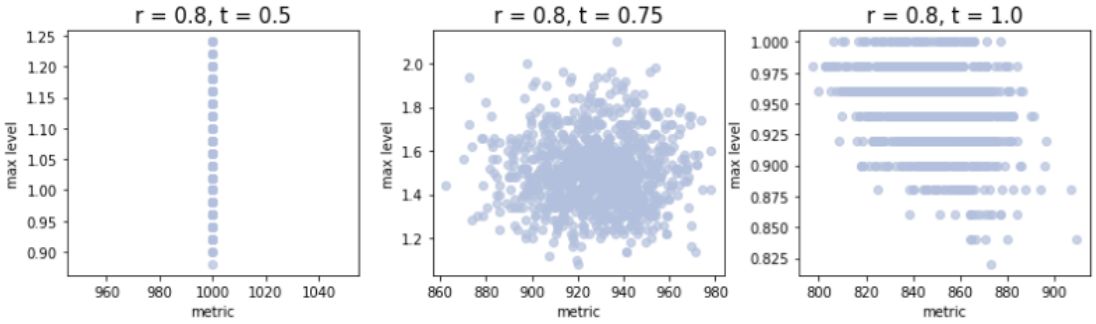


Figure 4.16: Behaviour of the metric and the maximum level parameters around the inflection point, shown for $n = 1000$.

connected regime. Hence, the representation of the relation between `metric` and `max_level` is provided for all these radii in Appendix B.3.

In Figure 4.12 we can see that the pattern *initial phase* - *critical phase* - *final phase* that has been already notice in all the other experiments as a characterization of the inflection point (or, inflection range). In the images generated for this experiment, it can be seen that for $r_n = 0.8$ and all values of n the unique-horizontal-line behaviour typical of the final phase in ERGs is again not reached. This is a common feature for the RGGs random social graphs with labeling function f_1 or f_2. It would be interesting to better investigate the theoretical reasons why this happens, but is for now left as a future possible investigation.

Chapter 5

Theoretical Results on ERG

In this section some of the evidences coming from the data collected through the experiments on ERGs are translated in properties of the LTR. Both the labeling functions defined in the experiments will be considered in the following properties, with the restriction of considering $t \in (0, 1]$: this because the case $t = 0$ is trivial, it makes the influence expansion reduce to a BFS visit of the graph. All the definitions used are given in Section 3.

We start reporting a tool, a simplified version of Chernoff bounds taken from document [14], used in the following proofs.

Proposition 5.1. *Let $X_1, \ldots, X_n$ independent random variables, Bernoulli distributed, each of them with probability parameter p_i.*
Let $\mu = \mathbb{E}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} p_i$. Then $\forall\ \delta \in (0, 1)$ we have:

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i < (1-\delta)\mu\Big) \leq \exp\Big(-\frac{\delta^2\mu}{2}\Big)\ ;$$
$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i > (1+\delta)\mu\Big) \leq \exp\Big(-\frac{\delta^2\mu}{3}\Big)\ .$$

5.1 Phase transition for the null maxlevel

As already pointed out in the analysis of the results done in Sections 4.2 and 4.3, the LTR metric on ERGs shows that for small probabilities the influence is not able to spread along the graph. This is represented by the fact that there are several nodes with a maximum expansion level equal to zero or one and with an average expansion level smaller than zero.

The property $\mathcal{P}_0$ that arises from the experimental data can be reformulated in the following way:

Definition 5.1. Let $i \in V$ be the seed of the influence expansion ($LTR(i)$). The $\mathcal{P}_0$ property is defined such that:

$$\mathcal{P}_0 \ : \text{"}\exists\ i \mid maxlevel(i) = 0\text{"}.$$

Remark 5.2. It is interesting to notice that the $\mathcal{P}_0$, being equivalent to asking

$$\mathcal{P}_0 \ : \text{"}\exists\ i \text{ s.t. the set} \{i\} \cup neigh(i) \text{ is a connected component of } G\text{"},$$

is independent from the labeling function. It is a property that only depends on the structure of the graph.

Theorem 5.3. *The property $\mathcal{P}_0$ has a sharp phase transition with respect the threshold sequence $a_n = \frac{\log n}{n}$ of the form*

$$\lim_{n\to+\infty} \mathbb{P}(\mathcal{P}_0 \textit{ holds for } (G(n,p_n)), f) = \begin{cases} 0 \ \textit{if } p_n \geq c\ \frac{\log n}{n},\ c > 1 \\ 1 \ \textit{if } p_n \leq c\ \frac{\log n}{n},\ c < 1. \end{cases}$$

where $G(n,p_n)$ is an undirected ERG graph. The phase transition is independent of the labeling function.

Proof. The proof of this theorem will distinguish between two cases, Part I and Part II.

Part I: Case $p_n \leq c\frac{\log n}{n},\ c < 1$

Under the hypothesis $p_n < c\frac{\log n}{n},\ c < 1$ we already know from Section 2.2 that the probability of having at least an isolated vertex goes to one for n going to infinity. Since picking an isolated vertex as seed of the LTR influence expansion will lead to have $maxlevel(i) = 0$ and every node of the graph is picked as seed, then the probability of $\mathcal{P}_0$ holding for $G(n,p_n)$ goes to one in this range of p_n's.

Part II: Case $p_n \geq c\frac{\log n}{n},\ c > 1$

Under the hypothesis $p_n \geq c\frac{\log n}{n},\ c > 1$ we are in the connected regime of the $G(n,p_n)$ graph. Here the graph is connected with probability one when n goes to infinity. Given the definition of the $maxlevel(i)$ random variable, if the initial activation set is connected to any other node not in the set the property cannot be true. This event actually takes place with probability one, so the property has asymptotically probability zero in this regime.

The only remaining case is when all the nodes in the network are in the initial activation set of i, i.e. they have at least a connection with i. This means that the

degree of i is supposed to be equal to $n-1$.
It is known that the following Simplified Chernoff bound holds (Prop. 5.3):

$$\mathbb{P}\bigg(d(i) = \sum_{j \neq i} \xi_{i,j} > (1+\delta)p_n(n-1)\bigg) \leq \exp\bigg(-\frac{\delta^2 p_n(n-1)}{3}\bigg) \tag{5.1}$$

for every $\delta \in (0,1)$, for all i nodes. For δ small and $p_n \geq c\frac{\log n}{n},\ c > 1$ this bound goes to zero for infinite n, so the claim follows.

□

Expected number of nodes with null maxlevel. We can derive a closed formula for the expected number of nodes s.t. $maxlevel(i) = 0$. We define a family of Bernoulli random variables Y_i : "$\{i\} \cup neigh(i)$ is a connected component of G", iid with probability of success

$$\begin{aligned}
\mathbb{P}(Y_i = 1) &= \sum_{k=0}^{n-1} \mathbb{P}(|neigh(i)| = k) \cdot \mathbb{P}(Y_i = 1 \mid |neigh(i)| = k) \\
&= \sum_{k=0}^{n-1} \binom{n-1}{k} p_n^k (1-p_n)^{(n-1-k)} \cdot (1-p_n)^{k(n-1-k)} \\
&= \sum_{k=0}^{n-1} \binom{n-1}{k} p_n^k (1-p_n)^{(k+1)(n-1-k)}
\end{aligned} \tag{5.2}$$

where $\mathbb{P}(|neigh(i)| = k) = \binom{n-1}{k} p_n^k (1-p_n)^{(n-1-k)}$ since we are choosing k nodes connected with i and the edges are independent Bernoulli r.v. of parameter p_n. Furthermore, $\mathbb{P}(Y_i = 1 \mid |neigh(i)| = k) = (1-p_n)^{k(n-1-k)}$ since we are excluding, for all the k nodes in $neigh(i)$, the possibility of having connections with the outside (the $n-1-k$ nodes remaining). The connections among the nodes inside $neigh(i)$ are not relevant for the probability, so we do not add any expression of their behaviour.

Then, the expected number of i with this property will be

$$\begin{aligned}
\mathbb{E}\bigg(\sum_{i=1}^{n} Y_i\bigg) &= \sum_{i=1}^{n} \mathbb{E}(Y_i) \\
&= \sum_{i=1}^{n} \mathbb{P}(Y_i = 1) \\
&= n \sum_{k=0}^{n-1} \binom{n-1}{k} p_n^k (1-p_n)^{(k+1)(n-1-k)} \ .
\end{aligned} \tag{5.3}$$

5.2 Significant labeling functions

The goal of a centrality measure is to compute the number of actors a node of the graphs can influence. Given this number, one can discriminate among nodes with a low power of influence and nodes with a high one. Labeling functions that only returns an LTR value equal to the initial activation set are considered not enough informative, i.e. not significant.

Definition 5.1. A labeling function f of a random social graph (G, f), $G = (V, E)$, is said to be **significant for** G with respect the LTM if the influence expansion process based on LTM is such that

$$\exists\ X = \{i\} \cup neigh(i) \subsetneq V \mid F(X) \neq X.$$

Remark 5.2. Since we are considering random graphs, the significance of a labeling function is not a deterministic concept, it is an event with an associated probability.

The Proposition 5.1 allows to determine values of the labeling functions that can lead to significant results for the LTR. In the case of LTR on ERG graph, the proposition can be reformulated as

Proposition 5.3. *Let i be a node of the random social graph $(G(n, p_n)), f)$ where $G(n, p_n)$ is a random variable corresponding to an ERG. Consider the iid Bernoulli random variables ξ_{ij}, $i \neq j$ of parameter p_n representing the edges involving i. Let $\mu_n := \mathbb{E}(\sum_{j\neq i}^{n} \xi_{ij}) = (n-1)p_n$. Then $\forall\ \delta \in (0, 1)$ we have:*

$$\mathbb{P}\left(d(i) = \sum_{j\neq i}^{n} \xi_{ij} < (1-\delta)\mu_n\right) \leq \exp\left(-\frac{\delta^2\mu_n}{2}\right);$$
$$\mathbb{P}\left(d(i) = \sum_{j\neq i}^{n} \xi_{ij} > (1+\delta)\mu_n\right) \leq \exp\left(-\frac{\delta^2\mu_n}{3}\right).$$

Stated all the instruments we need, we can prove the following results:

Lemma 5.4. *Let a_n be a real sequence defined as*

$$a_n = n \exp\left(-c(n-1)p_n\right)$$

with $c > 0$, p_n is a sequence of probabilities and $n \in \mathbb{N}$. Then, when n goes to infinity and $p_n^{-1} = o(\frac{n}{\log n})$ the sequence converges to zero.

Proof. Consider the a_n sequence formula: we know the asymptotic behaviour of the exponential function plus the fact that $c(n-1)p_n \geq 0$; we can study a_n's behaviour by analysing its natural logarithm.

$$\log a_n = \log n - c(n-1)p_n, \tag{5.4}$$

which goes to $-\infty$ when $p_n^{-1} = o(\frac{n}{\log n})$, proving the statement. □

Theorem 5.5. *Let $(G(n,p_n),f)$ be a random social graph where $G(n,p_n)$ is a random variable corresponding to an ERG and $f \in \{f_1,f_2\}$, as defined in Def. 3.2. Then $\exists\ \hat{n} \in \mathbb{N}$ such that $\forall\ n \geq \hat{n}$ the labeling function f is almost surely not significant for G when $t > p_n(1+\delta)k,\ \delta \in (0,1),\ k \geq 1$ and $p_n^{-1} = o(\frac{n}{\log n})$.*

Proof. Consider the Simplified Chernoff bounds for ERGs in Proposition 5.3 and a fixed node i of the graph. The estimation from above of the probability of the $d(i)$ random variable to assume values near its mean is given by exponential sequences. We know that:

$$0 \leq \mathbb{P}\bigg(d(i) < (1-\delta)\mu_n\bigg) \leq \exp\bigg(-\frac{\delta^2\mu_n}{2}\bigg)$$
$$0 \leq \mathbb{P}\bigg(d(i) > (1+\delta)\mu_n\bigg) \leq \exp\bigg(-\frac{\delta^2\mu_n}{3}\bigg)$$

for every $\delta \in (0,1)$, $\mu_n = (n-1)p_n$.
Now consider the event

$$A = \bigcap_{i=1}^{n} \{d(i) > (1+\delta)\mu_n\}. \tag{5.5}$$

Recalling basic properties of probability measures, we know that its probability can be rewritten as

$$\mathbb{P}(A) = 1 - \mathbb{P}\bigg(\bigcup_{i=1}^{n} \{d(i) > (1+\delta)\mu_n\}^C\bigg) \tag{5.6}$$
$$\geq 1 - \sum_{i=1}^{n} \mathbb{P}\bigg(\{d(i) > (1+\delta)\mu_n\}^C\bigg) \tag{5.7}$$
$$\geq 1 - n\exp\bigg(-\frac{\delta^2\mu_n}{3}\bigg). \tag{5.8}$$

The formulation of these inequalities meet the requirements of Lemma 5.4. Hence,

if we consider $p_n^{-1} = o(\frac{n}{\log n})$, we know that $\exists \hat{n} \in \mathbb{N}$ s.t. the event A holds almost surely $\forall\ n > \hat{n}$.
Given the LTM definition, it is clear that when

$$d(i) < f(i)\ \forall\ i \tag{5.9}$$

the only nodes that will contribute to the ranking value of $\hat{i}$ (i.e. the activated nodes) will be the ones in the initial activation set $\{\hat{i}\} \cup neigh(\hat{i})$. Given the estimation proved above, we can say that if

$$(1+\delta)(n-1)p_n < f(i) \tag{5.10}$$

for $n > \hat{n}$, the influence does not propagate outside the initial set almost surely.
Now consider the two labeling function used in this work. It can be noticed that

$$f_2(i) = t \cdot |neigh(i)| \leq t \cdot (n-1) = f_1(i). \tag{5.11}$$

Hence, using Equations (5.9) and (5.10), we can say that the values of the t parameter that give a not significant labeling function almost surely are the ones such that

$$t > p_n(1+\delta)k \tag{5.12}$$

when n is big enough, i.e. $n \geq \hat{n}$. The value of k will be exactly one when $f = f_1$, $k = \frac{n-1}{\min_{i \in V} |neigh(i)|}$ when $f = f_2$. The claim follows. □

This results gives a sort of explanation of the behaviour shown in the Phases 3 of Experiment 1 and 2 (Sections 4.2.3, 4.3.3) when the value of δ is small, even if the result obtained is proved for n sufficiently large.

5.3 Probability of extinction at level t

In Section 1.2 we have described the main features of the LTM influence expansion process. We have seen that the process dies out at most in $n = |V|$ steps. In this section we will analyse the probability of extinction of the process at level $t < n$.

We assume that $X = \{i\} \cup neigh(i)$ for some $i \in V$, according to the definition of LTR we have given in Section 1.3. The influence process stops at t when $F_{t-1}(X) = F_t(X)$. Since the $F_t(X)$ is computed for a random graph, it is a random variable with an associated probability of assuming a certain value. On $(G(n, p_n), f_1)$ ER influence graph, we have the extinction at level $t < n$ with probability

$$
\begin{aligned}
p_{ext}^t &= \mathbb{P}(|F_t - 1(X)| = |F_t(X)| \mid |F_{t-1}(X)| = y) \\
&= \begin{cases} \prod_{i \in F_{t-1}(X)^C} \sum_{k=0}^{f_1(i)-1} \binom{y}{k} p_n^k (1-p_n)^{y-k} & \text{if } f_1(i) - 1 \leq y \\ 1 & \text{otherwise} \end{cases} .
\end{aligned} \tag{5.13}
$$

The labeling function has been set as f_1 since it is constant on all nodes. Using the f_2, which does not have this property, makes impossible to express the p_{ext}^t in this way. In addition, notice that p_{ext}^t is not dependent on t, so $p_{ext}^t = p_{ext}$.

Since the labeling function is the $f_1(i) = \tau \cdot (n-1),\ \tau \in [0,1]$, constant on all nodes, we can rewrite the probability of extinction at level t knowing the $t-1$ state as

$$
p_{ext} = \begin{cases} \left(\sum_{k=0}^{\tau(n-1)-1} \binom{y}{k} p_n^k (1-p_n)^{y-k} \right)^{n-y} & \text{if } \tau \cdot (n-1) - 1 \leq y \\ 1 & \text{otherwise} \end{cases} . \tag{5.14}
$$

Consider the series on k without the exponent $n-y$. Let's assume that the variable y is dependent on n, $y = y_n \leq n$ since the process did not die yet, and $\tau \cdot (n-1) - 1 \leq y$ in order to have a meaningful binomial factor. We can rewrite in the following way:

$$
\begin{aligned}
a_n &= \sum_{k=0}^{\tau(n-1)-1} \binom{y_n}{k} p_n^k (1-p_n)^{y_n - k} \\
&= \sum_{k=0}^{y_n} \binom{y_n}{k} p_n^k (1-p_n)^{y_n-k} - \sum_{k=\tau(n-1)}^{y_n} \binom{y_n}{k} p_n^k (1-p_n)^{y_n-k} \\
&= 1 - \sum_{k=\tau(n-1)}^{y_n} \binom{y_n}{k} p_n^k (1-p_n)^{y_n-k} \\
&= 1 - a_n'
\end{aligned} \tag{5.15}
$$

since the first term is the normalized binomial series, known to be equal to 1 for every value of y_n, p_n. Hence, for the Cauchy convergence criterion we have that a_n' goes to zero. This means that the whole a_n sequence is convergent to 1 for n going to infinity.

Now we can analyse the possible behaviours of y_n with respect to n. We already know that if $\tau > p_n$ the f_1 labeling function is not significant for $G(n, p_n)$. This means that the process stops at level 1. If $\tau < p_n$, then we can check if $\tau \cdot (n-1) \leq y_n$. This condition is equivalent to $\tau \leq \frac{y_n+1}{n-1}$.

The possible cases are:

1. $\frac{y_n+1}{n-1} \to 0$. In this case, for n going to infinity, we do not have any τ that respects the condition $\tau \leq \frac{y_n+1}{n-1}$. The p_{ext} is one, so the process dies almost surely;

2. $\frac{y_n+1}{n-1} \to l$. For $\tau \leq \min\{l, \liminf_n p_n\}$ we have that the process could survive. The behaviour of p_{ext} has to be studied;

3. $\frac{y_n+1}{n-1} \to 1$. For all τ we have that the process could survive. The behaviour of p_{ext} has to be studied.

Let's consider the whole expression of p_{ext} when $\tau \leq \frac{y_n+1}{n-1}$, Equation (5.15). We have:

$$p_{ext} = a_n^{n-y_n} = \exp\Big((n-y_n)\log a_n\Big). \tag{5.16}$$

We can analyse the asymptotic behaviour of the argument of the exponential.

$$\begin{aligned}(n-y_n)\log a_n &= (n-y_n)\log(1-a'_n)\\ &\approx -(n-y_n)a'_n\end{aligned} \tag{5.17}$$

since we know that a'_n goes to zero. The quantity $n - y_n$ can only converge to a constant λ or positively diverge.

In the first case, which is equivalent to cases 2 and 3 of the list above (i.e. $y_n \asymp n$), $(n-y_n)\log a_n \to 0$ so $p_{ext} \to 1$. The process of extinction at level t happens almost surely for n going to infinity.

In the second case, which is equivalent to case 1 of the list above, we already know that the process extinction at level t happens almost surely for n going to infinity.

Hence, we can conclude that for n big the process will die at level $t < n$ almost surely no matter which asymptotic behaviour the active set size y_n have. As we already noticed, p_{ext} does not depend on t, so the result is true for all the levels. Then, we can say that the process stops at level 1 almost surely when n is big enough, independently of the behaviour of p_n.

Chapter 6

Conclusion

In this thesis we have analysed the behaviour of the LTR (FLTR in the directed case), an influence-based centrality measure, on random social graphs models. The study extends the previous works on real cases data, finding both similarities and differences.

Two labeling functions have been used in the experiments: the *neighbour threshold* and the *max neighbour threshold*.

The first one, defined as the percentage t of neighbours required to be active, is the classical definition that has been used in all the background studies.

The second one is the percentage t of the quantity $|V| - 1$, i.e. the maximum number of connections a node can have in $G = (V, E)$. It represents an approximation of the neighbour function, since it is not taking into account the actual realization of the random graph but it is considering the maximum value the degree random variable can assume.

Four different experiments have been performed, with different random graph models and labeling functions.

In experiments one and two has been performed the influence expansion on Erdös–Rényi Graphs $G(n, p_n)$. The main focus has been on the undirected definition of the graphs, since the connectivity phase transitions are proved only in this case. However, it has been observed that the directed case give the same results.

As regards the LTR with max neighbour threshold, it has shown an impossibility to let the influence spread outside the initial activation set and its immediate adjacent vertices when the probability p_n is not in the connected regime. This observation is translated into two properties, $\mathcal{P}_0$ and $\mathcal{P}_1$

Property 0.

$$\mathcal{P}_0 : \text{"}\exists i \mid \texttt{max\_level}(i) = 0\text{"}.$$

Property 1.

$$\mathcal{P}_1 : \text{"}\exists i \mid \texttt{max\_level}(i) \leq 1 \ \wedge \ \texttt{avg\_level}(i) \leq 0\text{"}.$$

that have empirically shown to have a sharp phase transition. Both of them seem to be independent of the percentage parameter that defines the labeling function. In addition, some interesting behaviors for the metric parameter itself are observed: the metric distribution on nodes is Gaussian when not trivial (i.e. only one value assumed); for high probability values in the connected regime, the metric shows to have an inflection point when the t percentage is increased. Around the inflection point the metric assumes more distinguished values and the maximum level parameter has a peak. In other words, there is an interval of t values for which the LTR centrality better characterized the ability of a vertex to influence the other nodes and this influence is better distributed in time. For the probabilities 0.6, 0.7, 0.8 we found respectively the inflection ranges [0.3, 0.4], [0.45, 0.55], [0.6, 0.7].

A scheme of the influence range evolution is here given:

1. an **initial phase** in which the metric value is still maximum but assumed after different time steps of the influence expansion;

2. a **critical phase** in which the algorithm finds more difficult to spread the influence (the metric value is not always maximum) but the maximum level reached has a peak;

3. a **final phase** in which the metric value is more stable and the maximum levels decrease again to one.

The LTR with neighbour threshold have given similar results below the connected regime: $\mathcal{P}_0$ and $\mathcal{P}_1$ have shown the same phase transition and are independent by the percentage parameter. Moreover, another property $\mathcal{P}_\in$ has been noticed but not further explored.

Property 2.

$$\mathcal{P}_2 : \text{"}\exists i \mid \texttt{metric}(i) = |GC|\text{"}.$$

Also, the observations on the metric parameter are similar: the distribution is still Gaussian; the inflection range is again observed with the same evolution with respect to the t parameter. The ranges are slightly different: [0.75,0.85] for $p_n = 0.8$, [0.85,0.75] for $p_n = 0.7$, [0.75,0.65] for $p_n = 0.6$.

Experiments three and four have preliminary analysed the LTR on Random Geometric Graphs. The simulations have been performed without refining the initial parameter definition, due to a lack of time. However, the results obtained are still enough to recognize a similarity between this analysis and the one performed on Binomial Graphs.

For the LTR with max neighbour threshold, the $\mathcal{P}_0$ and $\mathcal{P}_1$ properties were still observed and showed a phase transition. The first one still changes behavior when

the subcritical regime ends; the second one shows a threshold sequence with values in the connected regime. Again, these results do not change with the percentage parameter. Also, the inflection points show the same characteristics for the metric and the maximum level parameters, but we do not have enough data to determine an inflection range.

Regarding the LTR with neighbour threshold, the $\mathcal{P}_0$ showed the same phase transition. The $\mathcal{P}_1$ property showed the same behavior described for the previous threshold function apart for small values of the t parameter: in that case, the phase transition appears between the subcritical and the supercritical regimes. The $\mathcal{P}_2$ property is observed but not explored. The inflection point is still observed, but again the data are not enough to clearly observe where the inflection point starts and ends.

In general, the empirical results on the two models are similar: the only interesting difference the data generated for RGGs showed is that the $\mathcal{P}_1$ is dependent on the percentage parameter. Apart from that, we can affirm that the two labeling functions have a highly similar behavior. The features lost when using the approximated version are the $\mathcal{P}_2$ property (still not studied in depth) and the position of the inflection range (only determined on ERGs). It is important to underline that the running time is much lower for the experiments with the max neighbours thresholds with respect to the ones with the neighbour thresholds. A clear example is the running time for $n = 10^3$: with $f = f_1$ the numerical simulations were all finished after 3 hours, with $f = f_2$ each experiment with n and p_n fixed needed at least 1 hour to end.

The last part of the thesis is about a theoretical analysis of the problem studied: some properties arose from the experiments are formally proved for the ER model.

First, we have analysed and proved the phase transition for the $\mathcal{P}_0$ property. It is shown that the property is independent of the labeling function definition and that it has a sharp phase transition with respect to the sequence $\frac{\log n}{n}$. In addition, the expected number of nodes with this property has been computed.

The second part focuses on the labeling functions. We have given a definition of a significant labeling function for a graph. It has been used to give an explanation for the *final phase* for high values of n and $(G(n, p_n), f)$ ERG: it depends on the fact that the degree tends to assume values near the mean as the graphs size increases. Then, when $t > p_n$ it is impossible for the influence to spread outside the initial activation set.

The last one is about the probability of extinction of the process at level $t < n$: given an ER social graph $(G(n, p_n), f)$ and the number of influenced nodes y_n at the previous level, the probability that process stops is computed when $f = f_1$. Its asymptotic value for n big is computed depending on the behaviour of y_n: we have shown that in all the possible cases the probability of extinction goes to one

for every $t < n$ when n is big. In other words, we have proved that the probability of extinction at level 1 is asymptotically almost sure.

Some problems defined in this thesis can be better explored in future work. As already pointed out, the analysis of the $\mathcal{P}_2$ property can be an interesting study to perform, since it would characterize the metric on the Giant Component (supercritical) phase. In addition, the theoretical formulation and proof of the $\mathcal{P}_1$ and $\mathcal{P}_2$ phase transitions could be an interesting focus for future work, since it could have useful applications in network analysis.

Experiments three and four give preliminary results. All the simulations about the LTR on Random Geometric Graphs should be considered as a starting point for a deeper study, both empirical and theoretical.

Lastly, it would be interesting to perform a study similar to the one done in this thesis on other random network models. Some examples are the Barabási–Albert model and the Watts–Strogatz model (also known as small world model), more interesting since they can better represent the main features of the majority of real-world networks.

Bibliography

[1] F. Ajazi, *Random geometric graphs and their applications in neuronal modelling*, Doctoral Thesis, Lund University, Faculty of Science, Centre for Mathematical Sciences (2018).

[2] U. Alon, A. Drory, I. Balberg, *Systematic derivation of percolation thresholds in continuum systems*, Phys. Rev. A 42, 4634 (1990).

[3] U. Alon, I. Balberg, A. Drory, *New,heuristic, percolation criterion for continuum systems*, Phys. Rev. Lett. 66, 2879 (1991).

[4] E. Bakshy, J. M. Hofman, W. A. Mason, D. J. Watts, *Everyone's an influencer: Quantifying influence on twitter*, In Proc. of the Fourth ACM Int. Conf. on Web Search and Data Mining, WSDM '11, ACM (2011), 65–74.

[5] J. R. Banavar, A. Maritan, A. Rinaldo, *Size and form in efficient transportation networks*, Nature, London, 399, 130 (1999).

[6] A.L. Barabási, M. Pósfai, *Network science*. Cambridge: Cambridge University Press (2016).

[7] M. Barthelemy, *Spatial Networks*, Physics Reports 499 (2011), 1–101.

[8] J. Cannarella, J.A. Spechler, *Epidemiological modeling of online social network dynamics*, arXiv (2014), arXiv:1401.4208.

[9] W. Chen, L.V. Lakshmanan, C. Castillo, *Information and influence propagation in social networks* Morgan and Claypool, San Rafael (2013).

[10] W. Chen, C. Wang,Y. Wang, *Scalable influence maximization for prevalent viral marketing in large-scale social networks*, In Proc. of the 16th ACM SIGKDDInt. Conf. on Knowledge Discovery and Data Mining (2010).

[11] D.J. Daley, D.G. Kendall *Epidemics and rumors Nature*, 204 (4963) (1964), p. 1118.

[12] M. De Domenico, A. Lima, P. Mougel, M. Musolesi, *The Anatomy of a Scientific Rumor, Higgs data set*. In: Scientific Reports 3 2980 (2013).

[13] J. Díaz, D. Mitsche, X. Perez, *Dynamic Random Geometric Graphs*, In Proc. of the ACM SODA'07, New Orleans, LA (2007).

[14] J. Díaz, J. Petit and M. Serna, *A guide to concentration bounds*, S. Rajasekaram, Pet al (eds.), Handbook of Randomized Computing, Volume 2, pp, 457–507 (2001).

[15] X.J. Ding, *Research on propagation model of public opinion topics based on SCIR in microblogging*, Comput. Eng. Appl. (2015) 51, 20–26.

[16] A. Domínguez Besserer, *Estudio de mecanismos de ponderación de influencia y su efecto en el Forward Linear Threshold Rank*, TFG, Facultat d'Informàtica de Barcelona (2020).

[17] P. Erdös, A. Rényi, *On Random Graphs. I*, Publicationes Mathematicae. 6 (1959), 290–297.

[18] P. Erdös, A. Rényi, *On the evolution of random graphs*, Publication of the Mathematical Institute of the Hungarian Academy of Sciences, vol. 5 (1960), 17–61.

[19] L. Feng, Y. Hu, B. Li, H.E. Stanley, S. Havlin, L.A. Braunstein, *Competing for attention in social media under information overload conditions*, PLoS ONE (2015), 10, e0126090.

[20] L. Freeman, *A set of measures of centrality based on betweenness*. Sociometry 40 (1977), 35–41.

[21] A. Frieze, M. Karonski, *Introduction to Random Graphs*, Cambridge University Press (2016).

[22] D. Gao, *Opinion influence and diffusion in social network*. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information (2012) Aug 12–16; Portland, OR, USA; (2012) p. 997.

[23] P. García Rodríguez, *Estudio experimental del Forward Linear Threshold Rank*, TFG, Facultat d'Informàtica de Barcelona (2019).

[24] E.N. Gilbert, *Random Graphs*, Volume 30, Number 4 (1959), 1141–1144.

[25] E.N. Gilbert, *Random Plane Networks*, Journal of the Society for Industrial and Applied Mathematics 9 (4) (1961), 533–543.

[26] J. Goldenberg, B. Libai, E. Muller, *Using complex systems analysis to advance marketing theory development*, Tech. rep., Academy of Marketing Science Review (2001).

[27] G. Golnari, A. Asiaee, A. Banerjee, Z.L. Zhang, *Revisiting non-progressive influence models: Scalable influence maximization*. In: Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (2015) Jul 12–16; Amsterdam, The Netherlands (2015).

[28] M. Granovetter, *Threshold models of collective behavior*, Am. J. Sociol. 83 (6) (1978), 1420–1443.

[29] D. Kempe, J.M. Kleinberg, É. Tardos, *Influential nodes in a diffusion model for social networks*, in: L. Caires, G.F. Italiano, L. Monteiro, C. Palamidessi, M. Yung (Eds.), Automata, Languages and Programming, 32nd International Colloquium, ICALP 2005, Lisbon, Portugal, July 11–15 2005, Proceedings, Lecture Notes in Computer Science, 3580, Springer (2005), 1127–1138.

[30] D. Kempe, J.M. Kleinberg, É. Tardos. *Maximizing the spread of influence through a social network*, In Proc. of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2003), 137–146.

[31] M. Krivelevich, B. Sudakov, *The Phase Transition in Random Graphs: A Simple Proof*, Random Structures and Algorithms 43 (2013).

[32] J. Leskovec, D. Huttenlocher, J. Kleinberg, *Predicting positive and negative links in online social networks*,In Proc. of the 19th international conference on World wide web (WWW '10). Association for Computing Machinery, New York, NY, USA (2010), 641–650.

[33] L. Kan, Z. Lin , H. Heyan, *Social Influence Analysis: Models, Methods, and Evaluation.*, Engineering 4 (1) (2018), 40 –46.

[34] S.C. Lin, S.D. Lin, M.S. Chen, *A learning-based framework to handle multi-round multi-party influence maximization on social networks*. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015) Aug 10–13; Sydney, Australia (2015) p. 695–704.

[35] R. Lyons, Y. Peres, *Probability on Trees and Networks*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, New York (2016).

[36] M. E. J. Newman, *Spread of epidemic disease on networks*, Phys. Rev. E, 66:016128 (2002).

[37] L. Page, S. Brin, R. Motwani, T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report, Stanford Digital Library (1999).

[38] R. Pastorsatorras, *Epidemic spreading in scale-free networks*, Phys. Rev. Lett. (2001) 86, 3200–3203.

[39] M. Penrose, *On the spread-out limit for bond and continuum percolation.* Ann. Appl. Probab. 3 (1993), 253–276.

[40] M. Penrose, *Random Geometric Graphs*, Oxford studies in probability, Oxford University Press (2003).

[41] J. Quantanilla, S. Torquato, R.M. Ziff, *Efficient measurements of the percolation threshold for fully penetrable disks*, J. Phys. A 33, L399 (2000).

[42] F. Riquelme, P. Gonzalez-Cantergiani, X. Molinero, M. Serna. *Centrality measure in social networks based on linear threshold model*, Knowledge-Based Systems 140 (2018), 92–102.

[43] F. Riquelme, P. Gonzalez-Cantergiani, X. Molinero, M. Serna. *The neighborhood role in the linear threshold rank on social networks*, Physica A: Statistical Mechanics and its Applications 528 (2019).

[44] T. Schelling, *Micromotives and Macrobehavior*, Norton (1978).

[45] E. Van Hove, *The linear threshold rank as centrality measure in social networks*, Internship report, Facultat d'Informàtica de Barcelona (2018).

[46] X. Wang, J. Jia, J. Tang, B. Wu, L. Cai, L. Xie, *Modeling emotion influence in image social networks IEEE Trans Affect Comp*, 6 (3) (2015) pp. 286–297.

[47] C. Wang, X.y. Yang, K. Xu, J.F. Ma, *Seir-based model for the information spreading over SNS*, Tien Tzu Hsueh Pao/Acta Electron. Sin. (2014), 42, 2325–2330.

[48] Q. Wang, Z. Lin, Y. Jin, S. Cheng, T. Yang, *Esis: Emotion-based spreader–ignorant–stifler model for information diffusion*, Knowl.-Based Syst. (2015) 81, 46–55.

[49] R. Xu, H.Li, C. Xing, *Research on information dissemination model for social networking services*, Int. J. Comput. Sci. Appl. (2013) 2, 1–6.

Appendix A

Images for Erdös–Rényi Graphs

A.1 LTR with max neighbour threshold

Outputs per node computed as the average on all the realizations, here only shown for the initial probability and t sets.

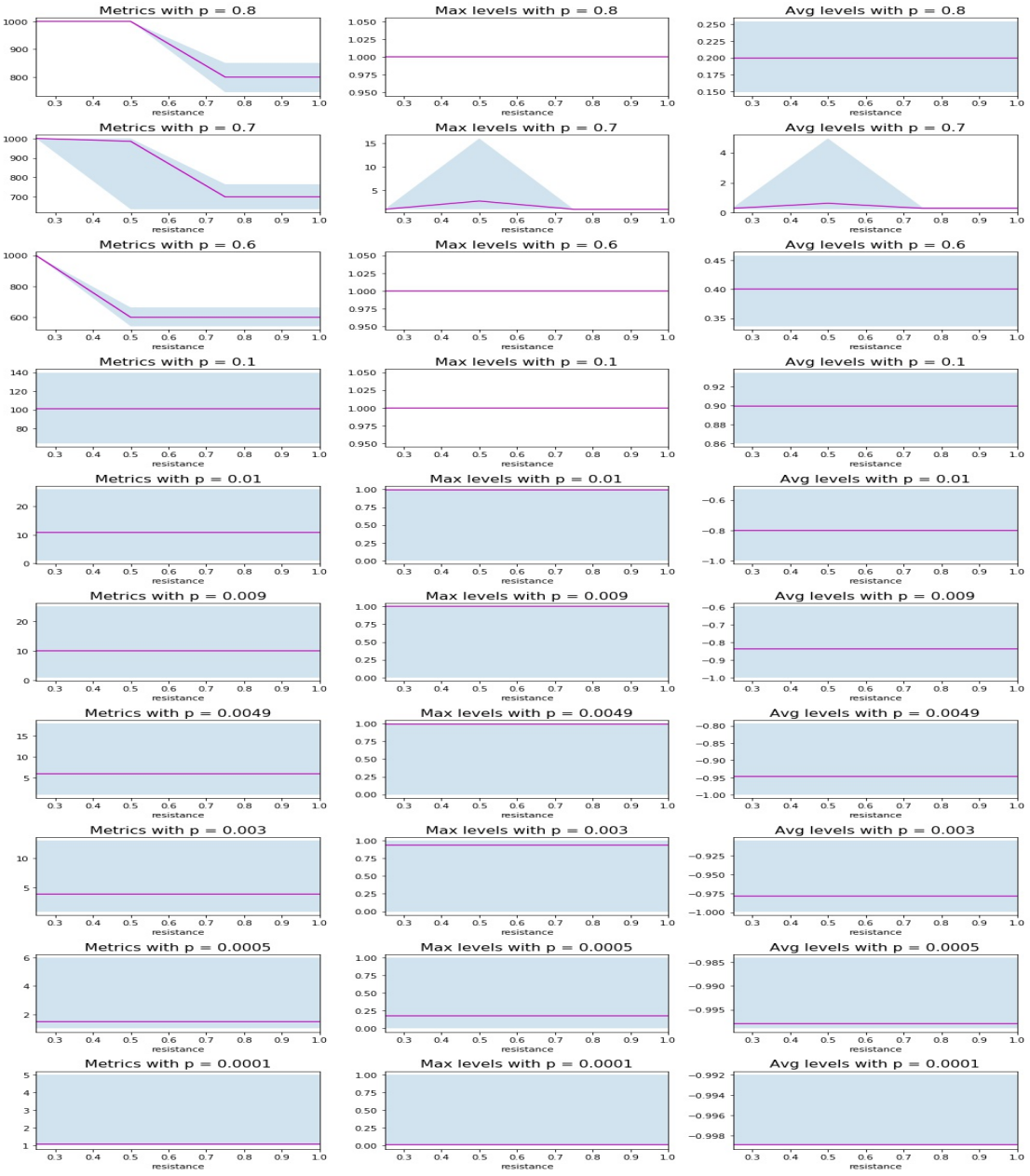


Figure A.1: $n = 1000$, undirected. Violet line: mean value, Blue area: [min, max] range.

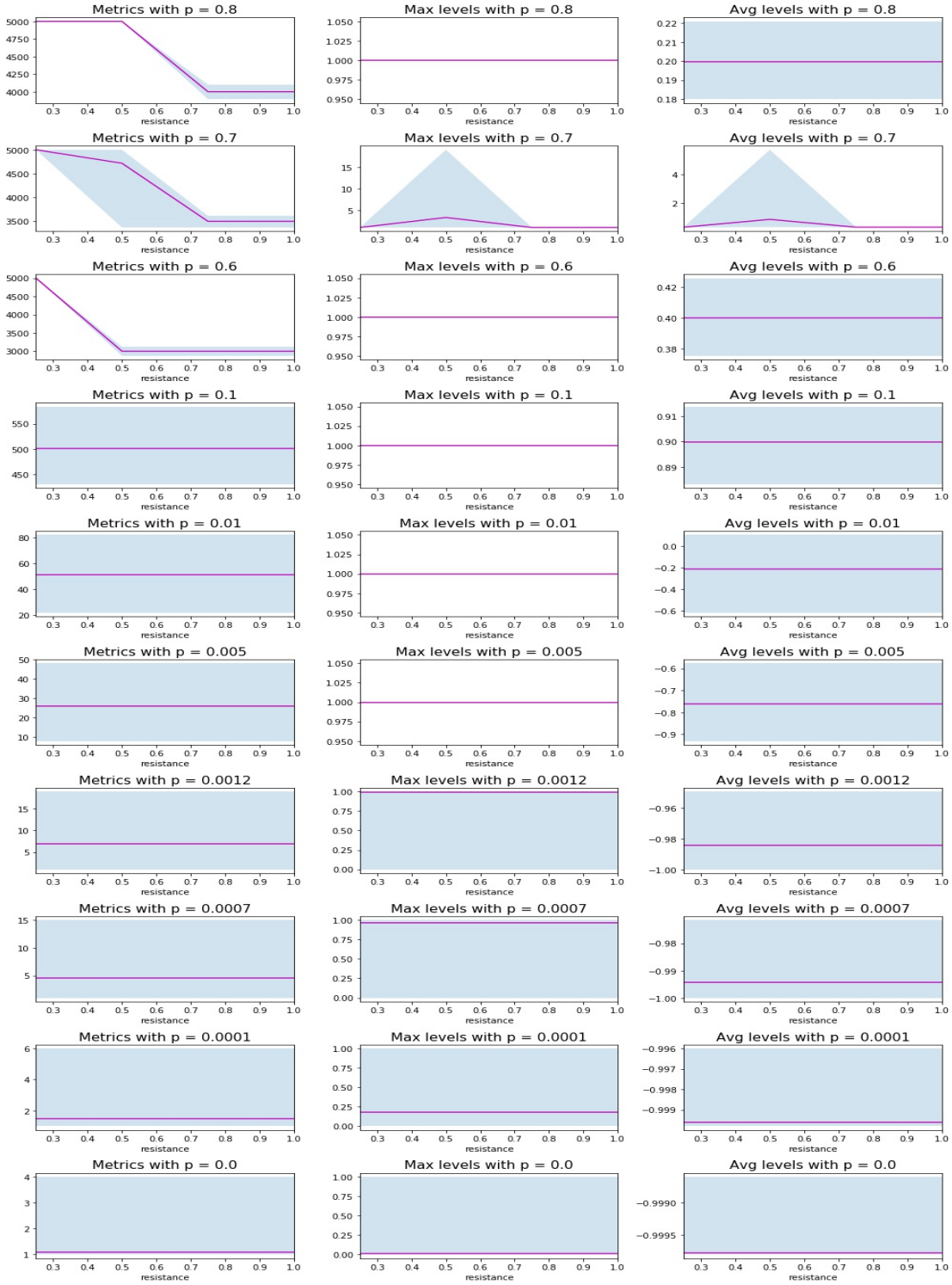


Figure A.2: $n = 5000$, undirected. Violet line: mean value, Blue area: [min, max] range.

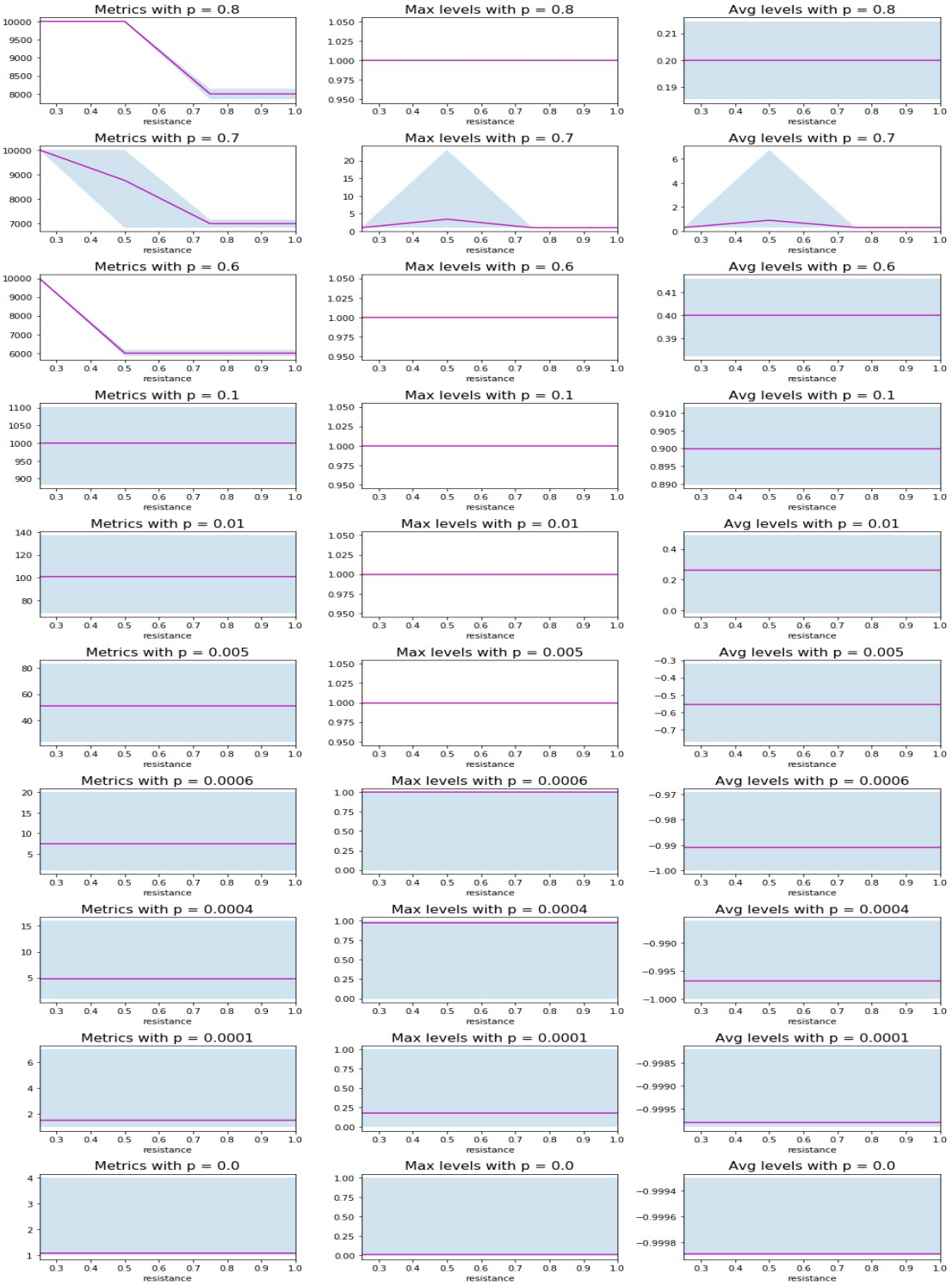


Figure A.3: $n = 10000$, undirected. Violet line: mean value, Blue area: [min, max] range.

A.1.1 Distribution the LTR with max neighbour threshold

Here the approximated density of the `metric` parameter is given, plus a normality test (QQplot). x axis: $\mathbb{E}_{i,k}$(`metric`); y axis: probability.

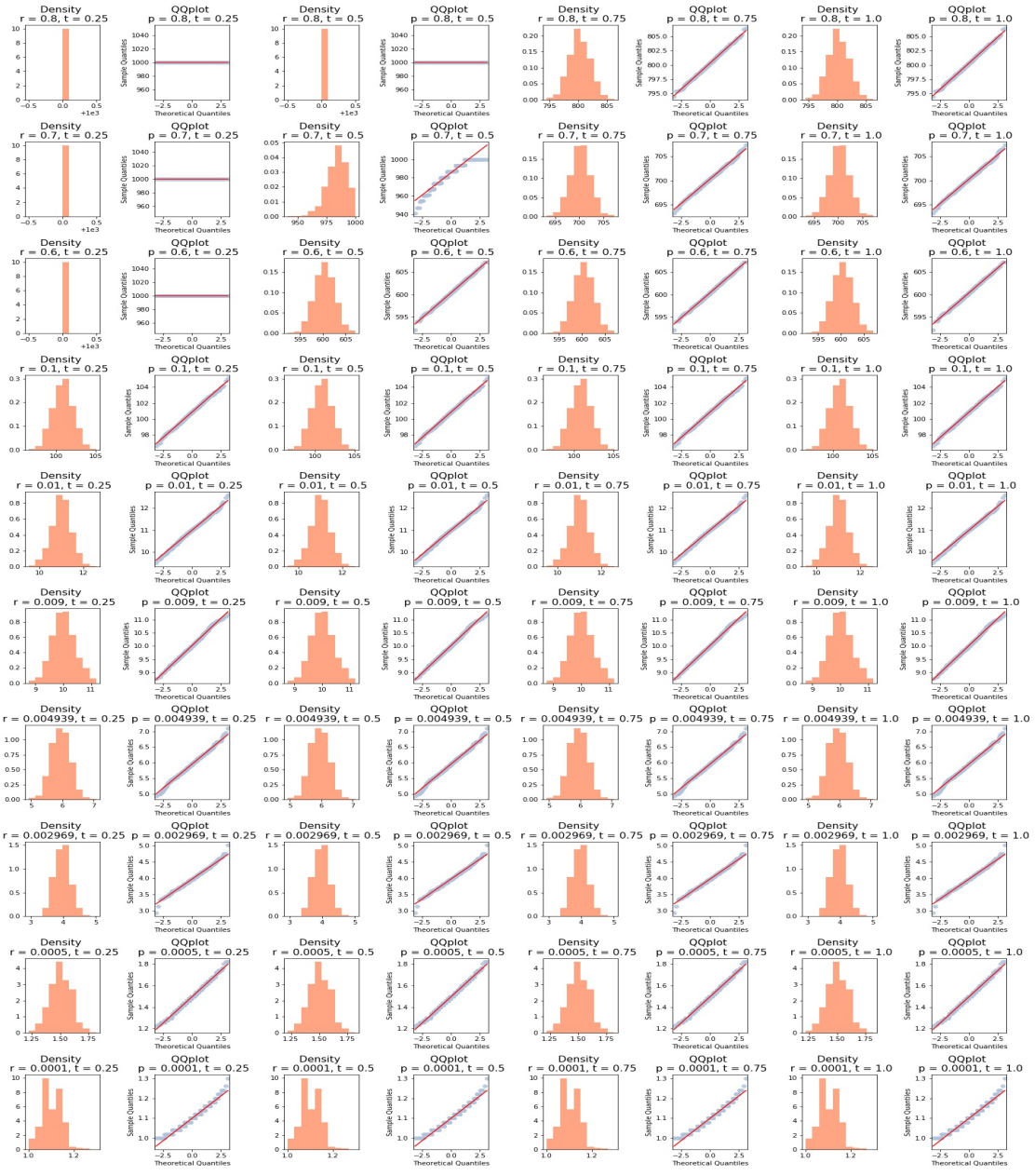


Figure A.4: $n = 1000$, undirected.

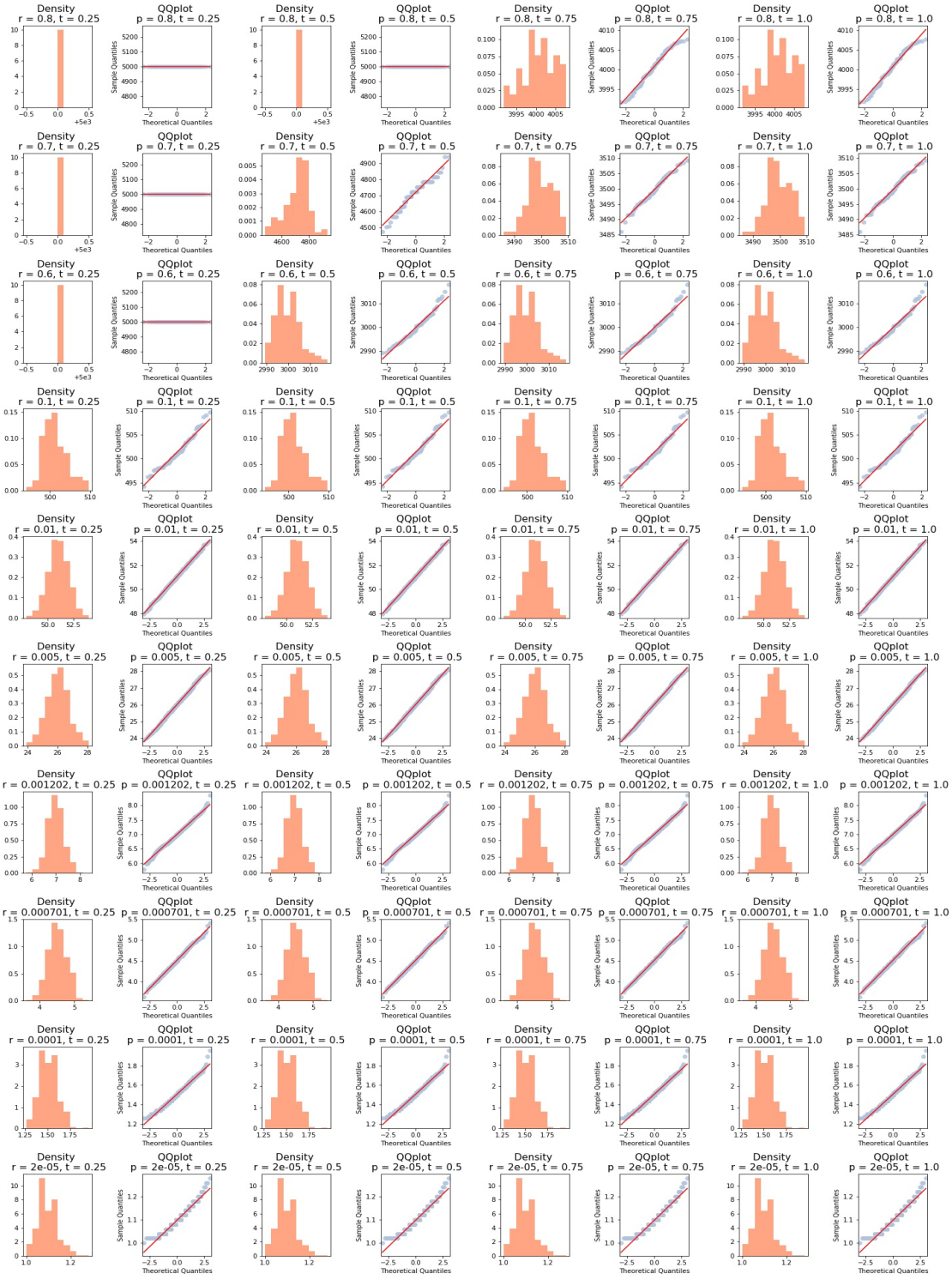


Figure A.5: $n = 5000$, undirected.

Figure A.6: $n = 10000$, undirected.

A.1.2 Refinement on high probabilities: inflection point.

Here represented the results for the refined t parameter (according to First experiment phase 3, Section 4.2) and sample size 100.

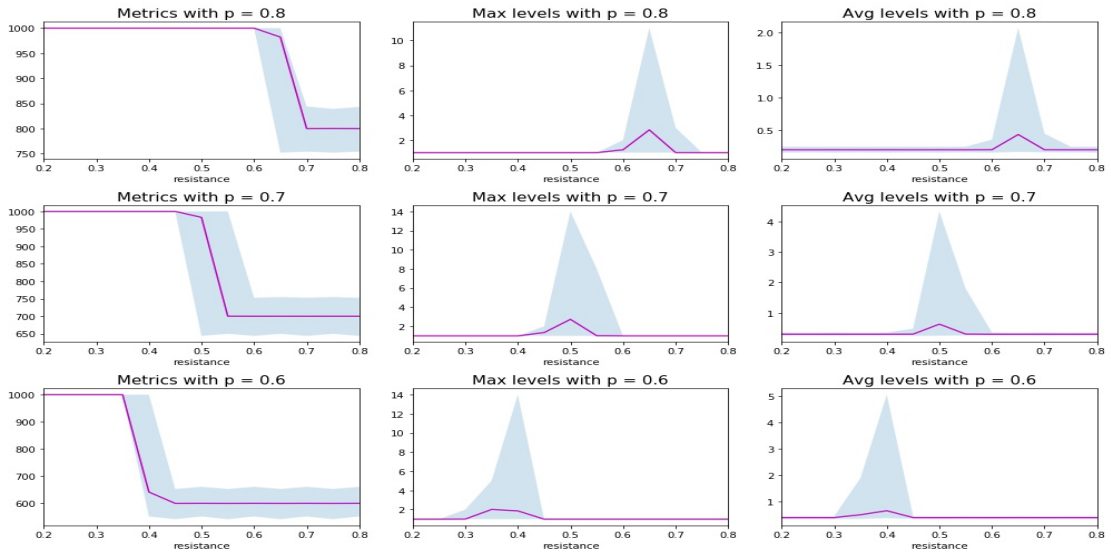


Figure A.7: $n = 1000$, undirected. Violet line: mean value, Blue area: [min, max] range.

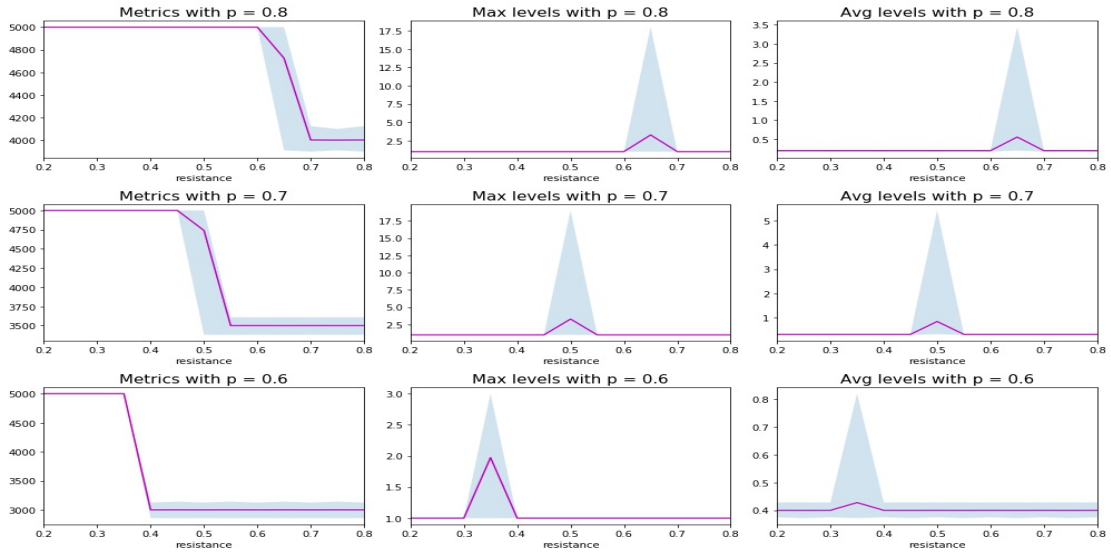


Figure A.8: $n = 5000$, undirected. Violet line: mean value, Blue area: [min, max] range.

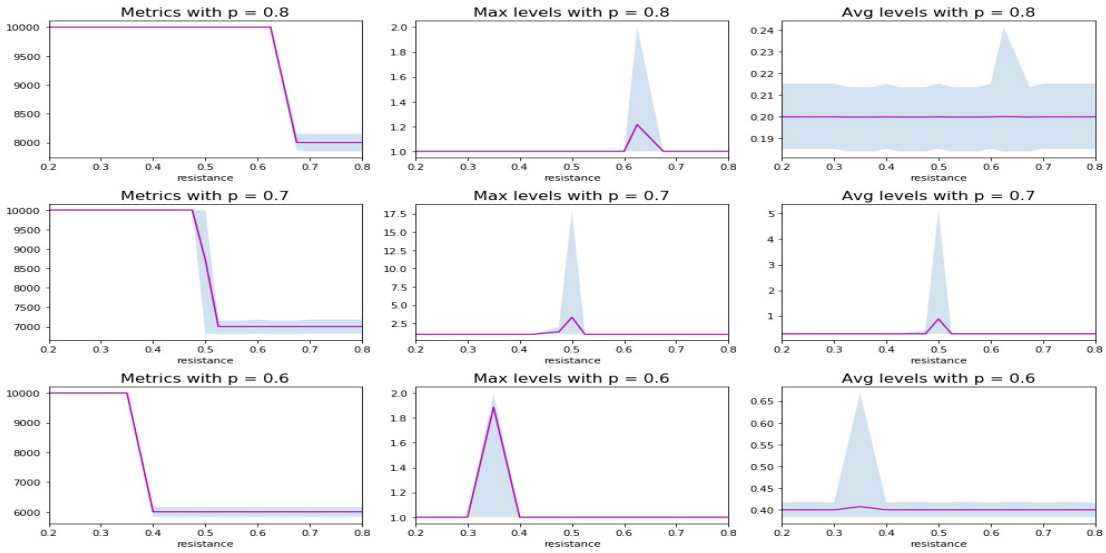


Figure A.9: $n = 10000$, undirected. Violet line: mean value, Blue area: [min, max] range.

A representation of the behaviour of the `metric` versus the `max_level` parameters is here provided to better characterize the inflection point.

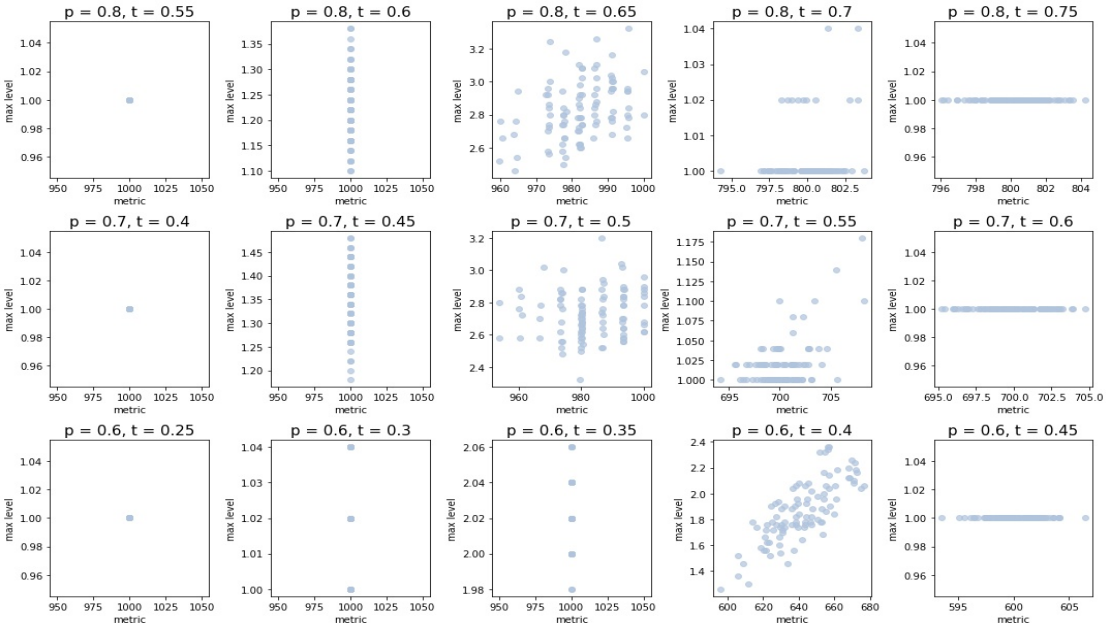


Figure A.10: $n = 1000$, undirected.

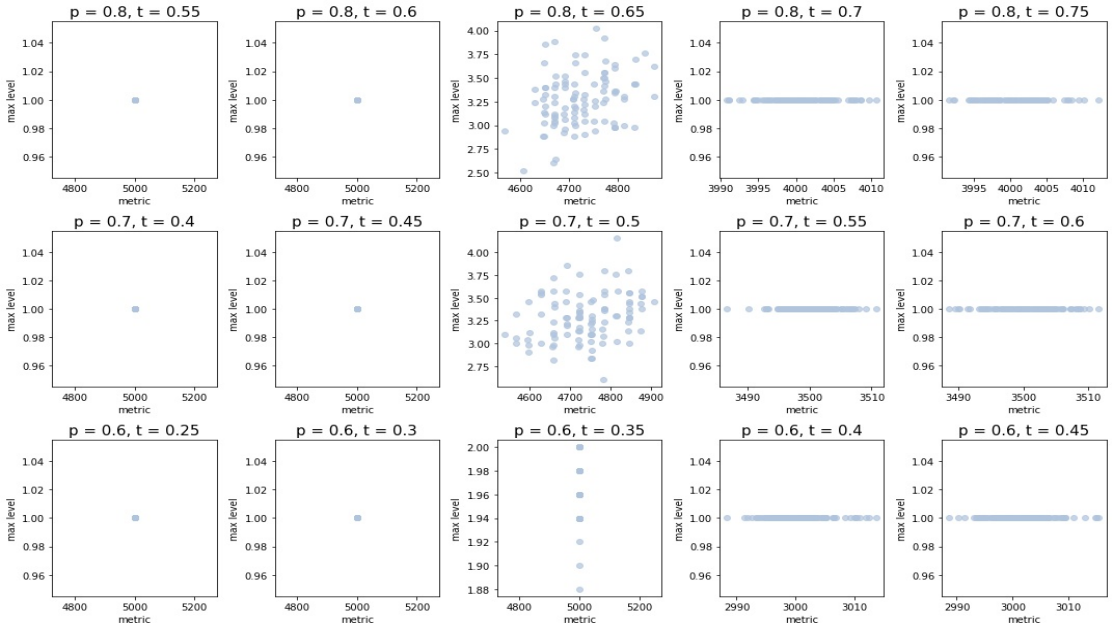


Figure A.11: $n = 5000$, undirected.

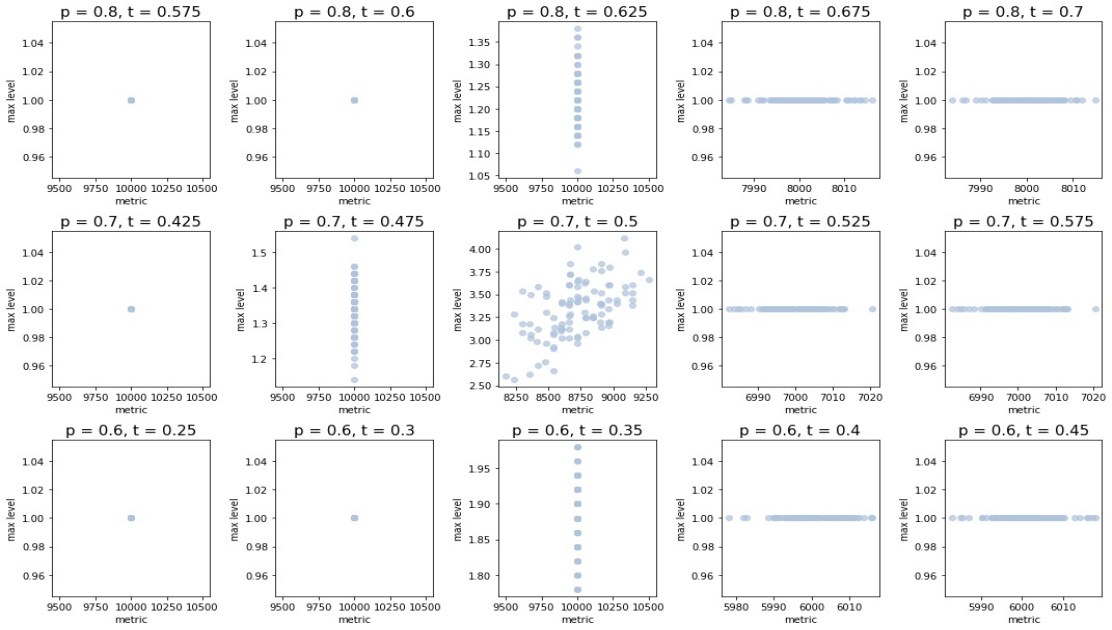
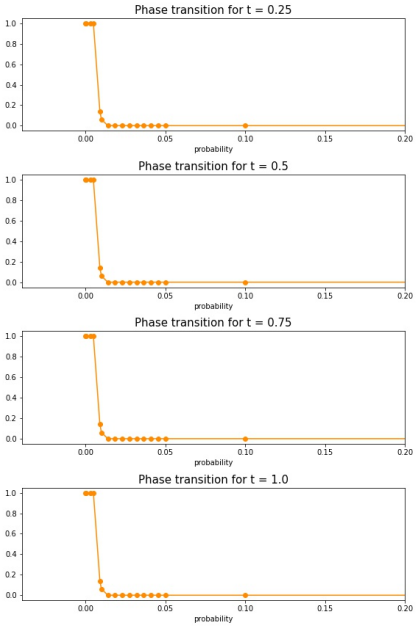


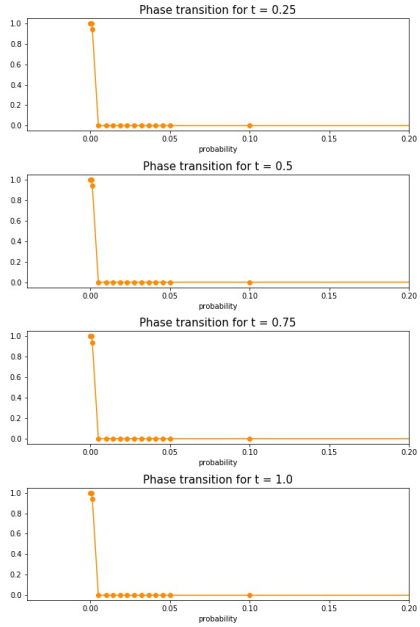
Figure A.12: $n = 10000$, undirected.

A.2 Phase transitions with max neighbour threshold

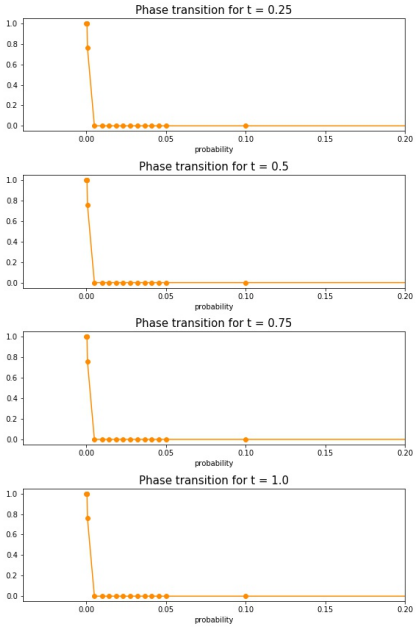
A.2.1 $\mathcal{P}_0$ phase transitions with max neighbour threshold



(a) $n = 1000$, undirected.



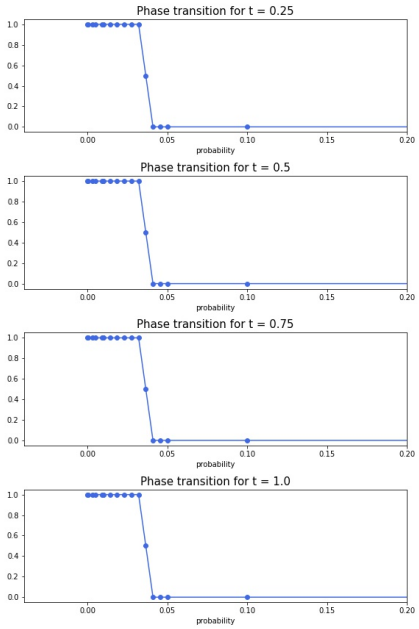
(b) $n = 5000$, undirected.



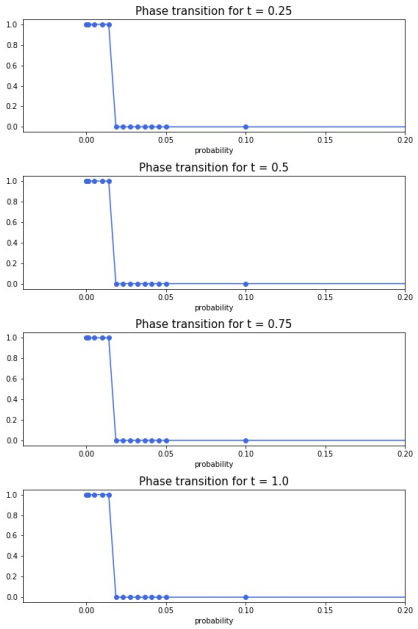
(c) $n = 10000$, undirected.

Figure A.13: Representation of the mean truth value assumed by the $\mathcal{P}_0$ property, mean computed on all the realizations.

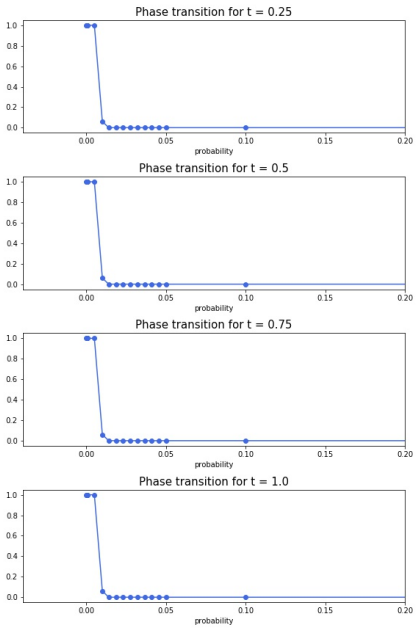
A.2.2 $\mathcal{P}_1$ phase transitions with max neighbour threshold



(a) $n = 1000$, undirected.



(b) $n = 5000$, undirected.



(c) $n = 10000$, undirected.

Figure A.14: Representation of the mean truth value assumed by the $\mathcal{P}_1$ property, mean computed on all the realizations.

A.2.3 Threshold sequences approximations

Representation of the last/first value of p_n for which the considered property was true/false in every realization. Here are shown the results only for the undirected case.

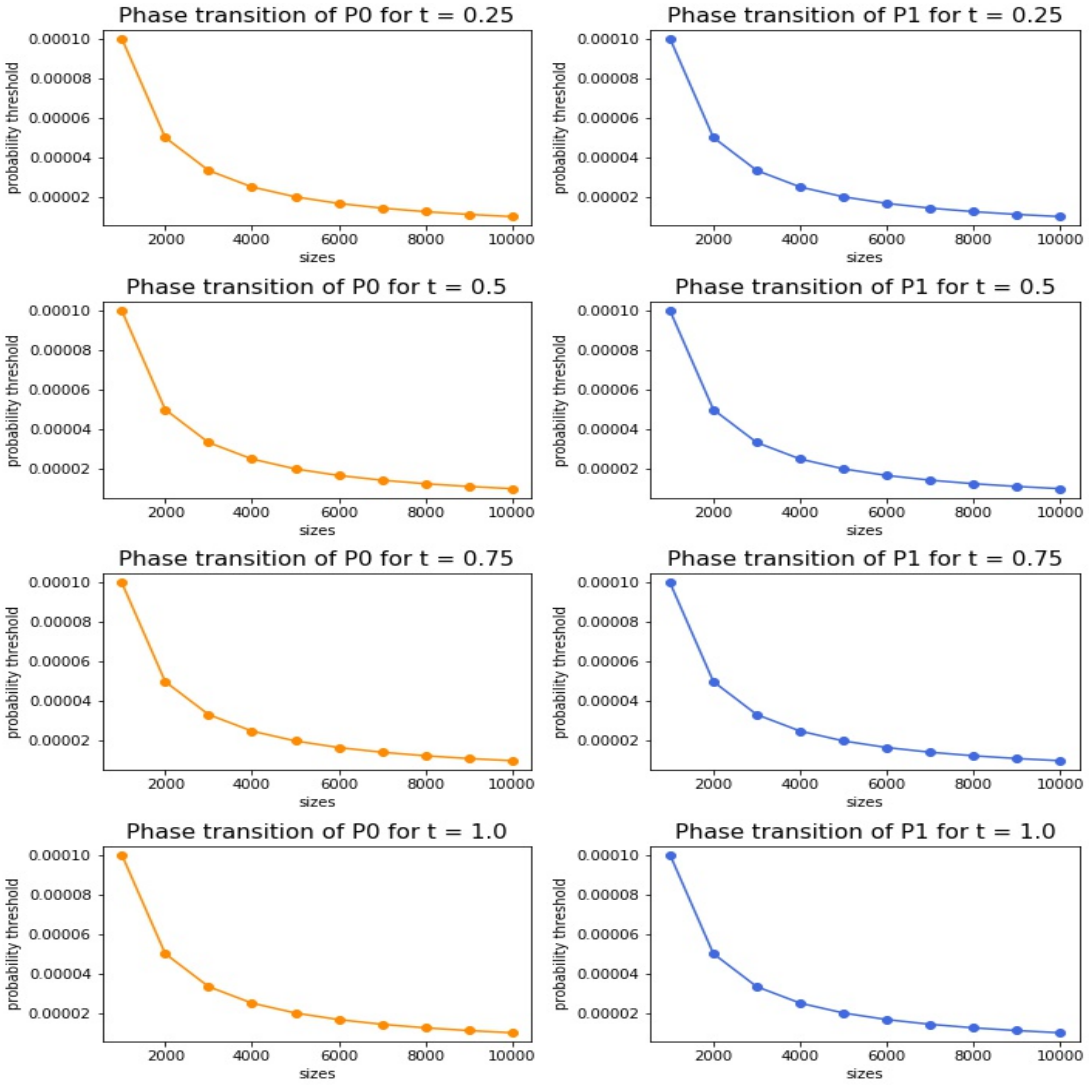


Figure A.15: undirected graphs, last probability for which the properties were always true.

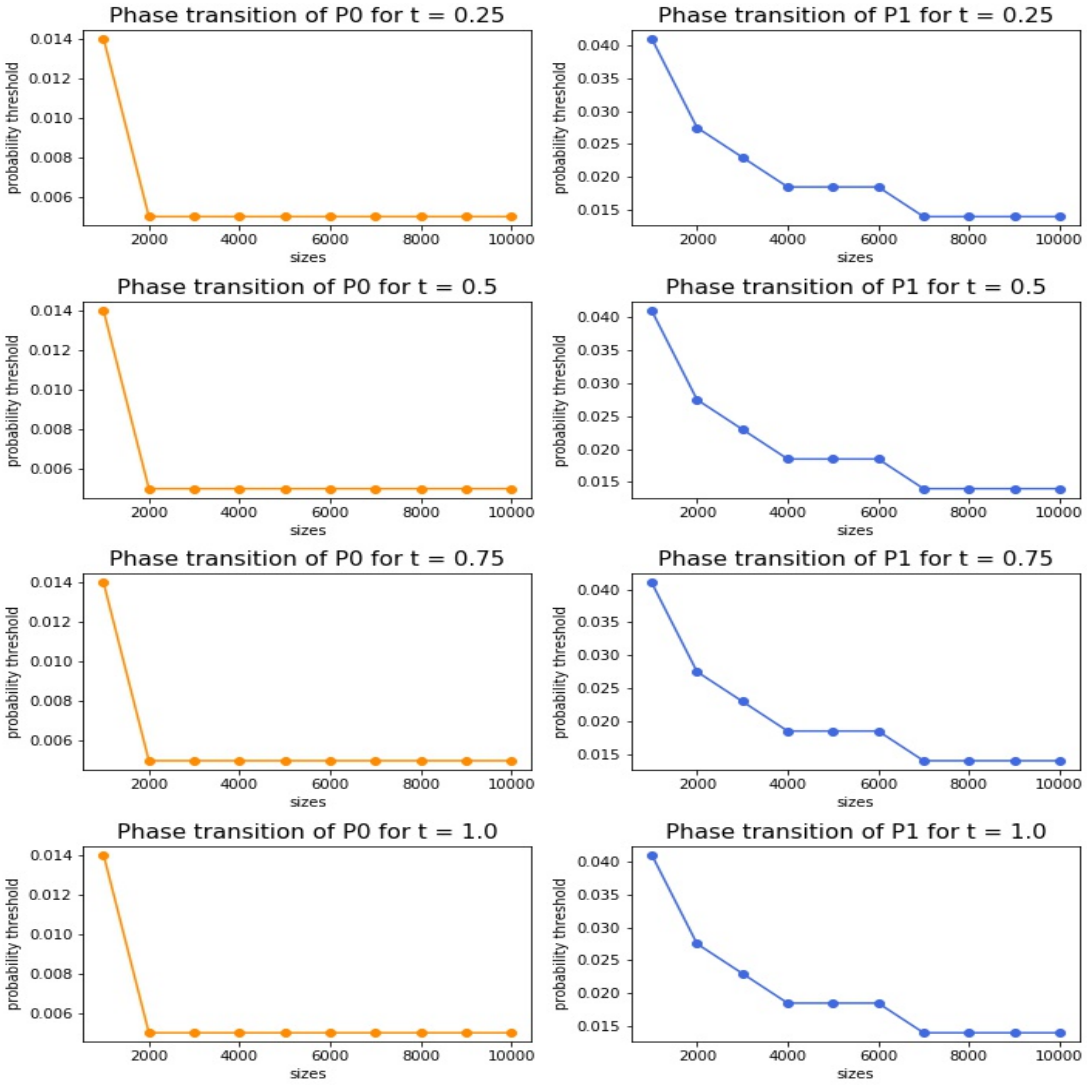


Figure A.16: undirected graphs, first probability for which the properties were always false.

A.3 LTR with neighbour threshold

Outputs per node computed as the average on all the realizations, here only shown for the initial probability and t sets.

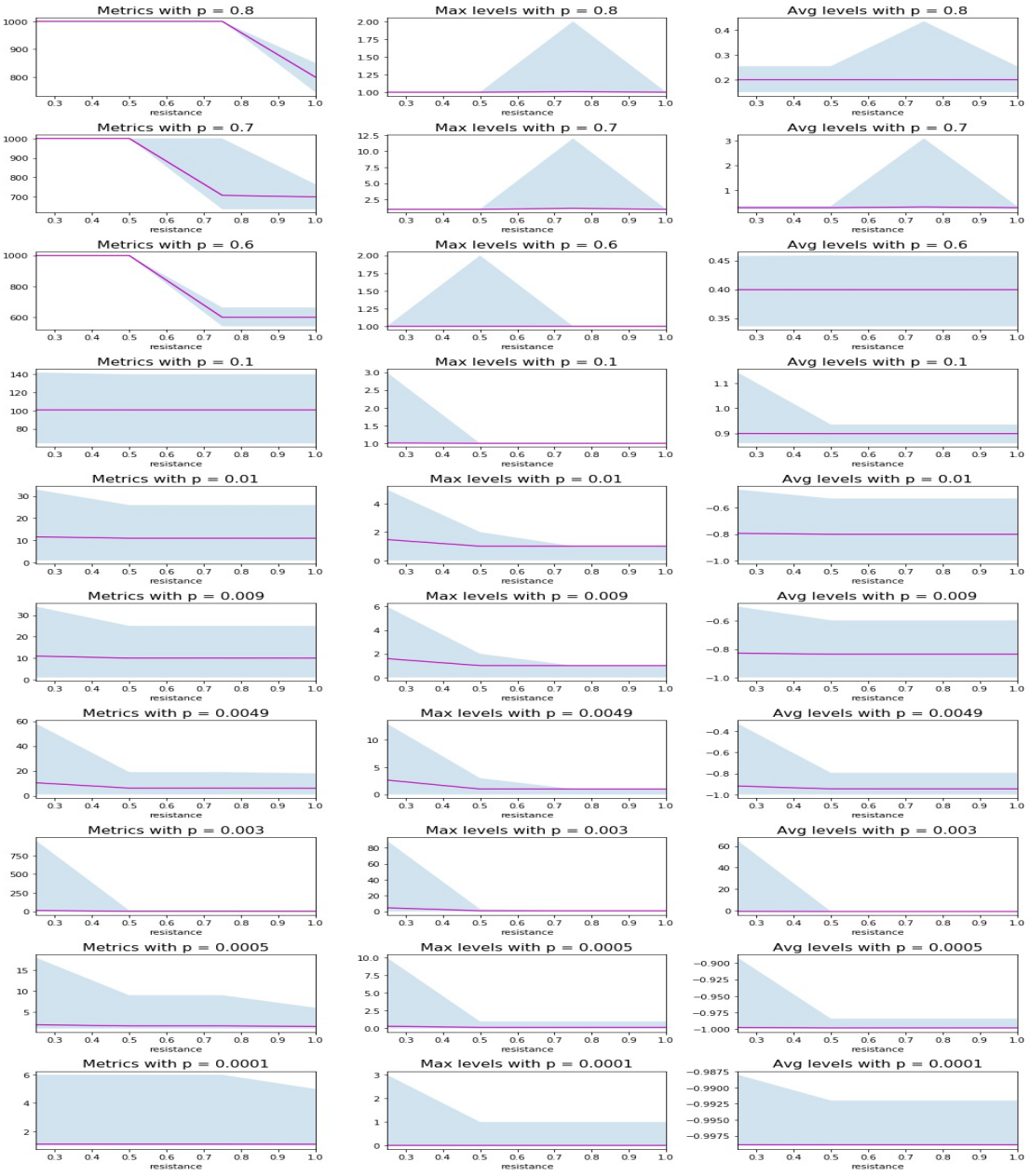


Figure A.17: $n = 1000$, undirected. Violet line: mean value, Blue area: [min, max] range.

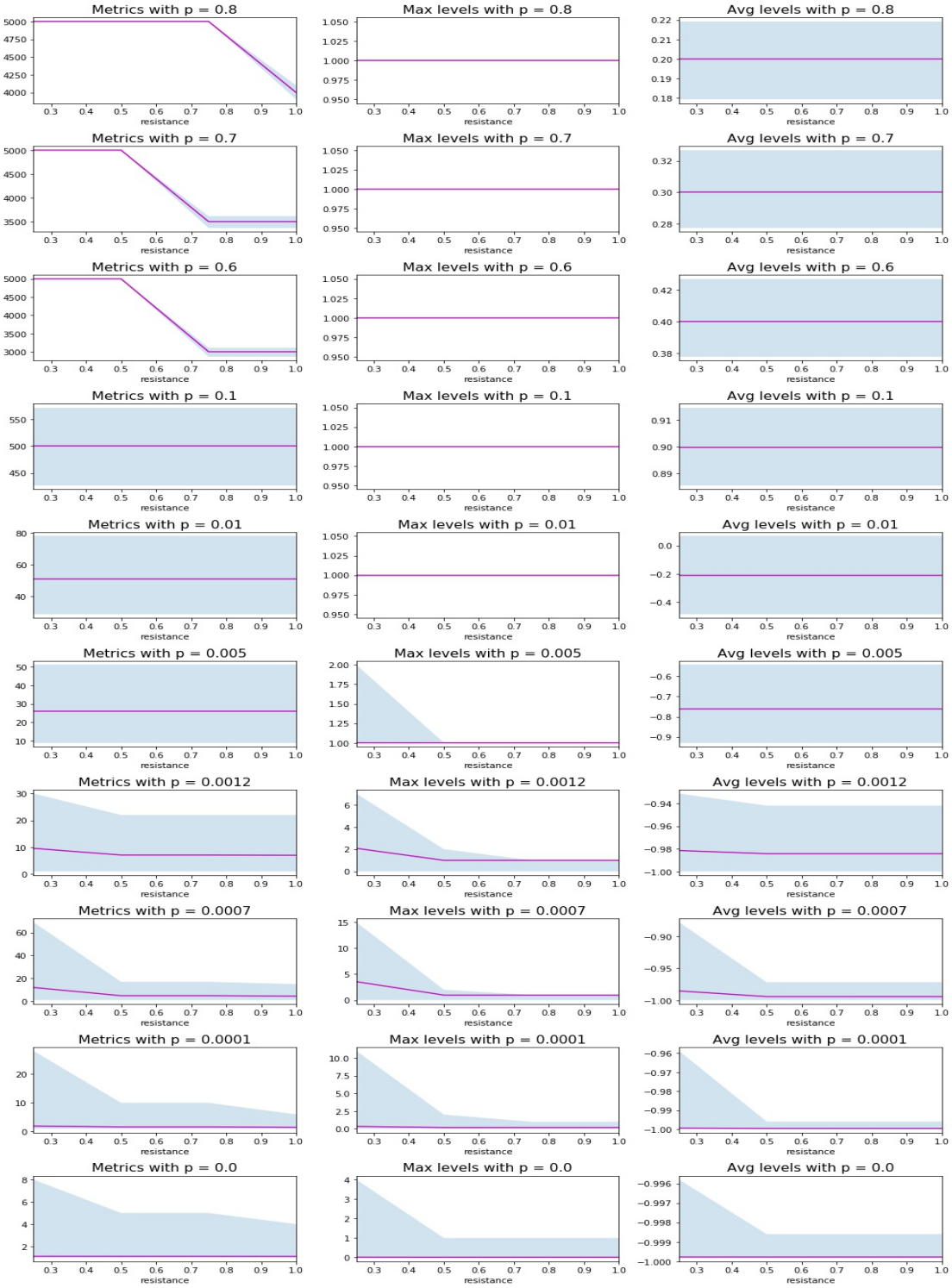


Figure A.18: $n = 5000$, undirected. Violet line: mean value, Blue area: [min, max] range.

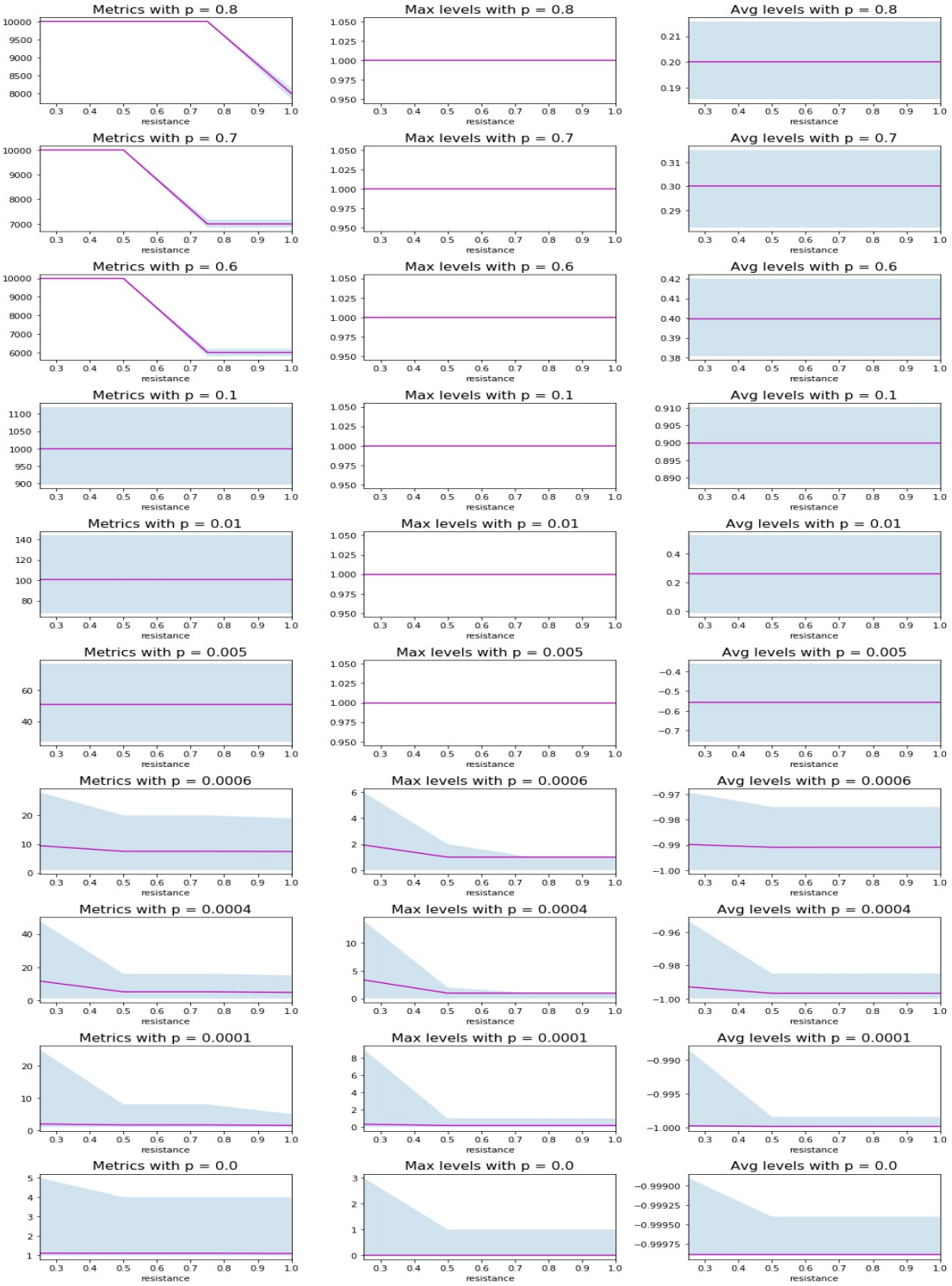


Figure A.19: $n = 10000$, undirected. Violet line: mean value, Blue area: [min, max] range.

A.3.1 Distribution of the LTR with neighbour threshold

Here the approximated density of the `metric` parameter is given, plus a normality test (QQplot). x axis: $\mathbb{E}_{i,k}(\texttt{metric})$; y axis: probability.

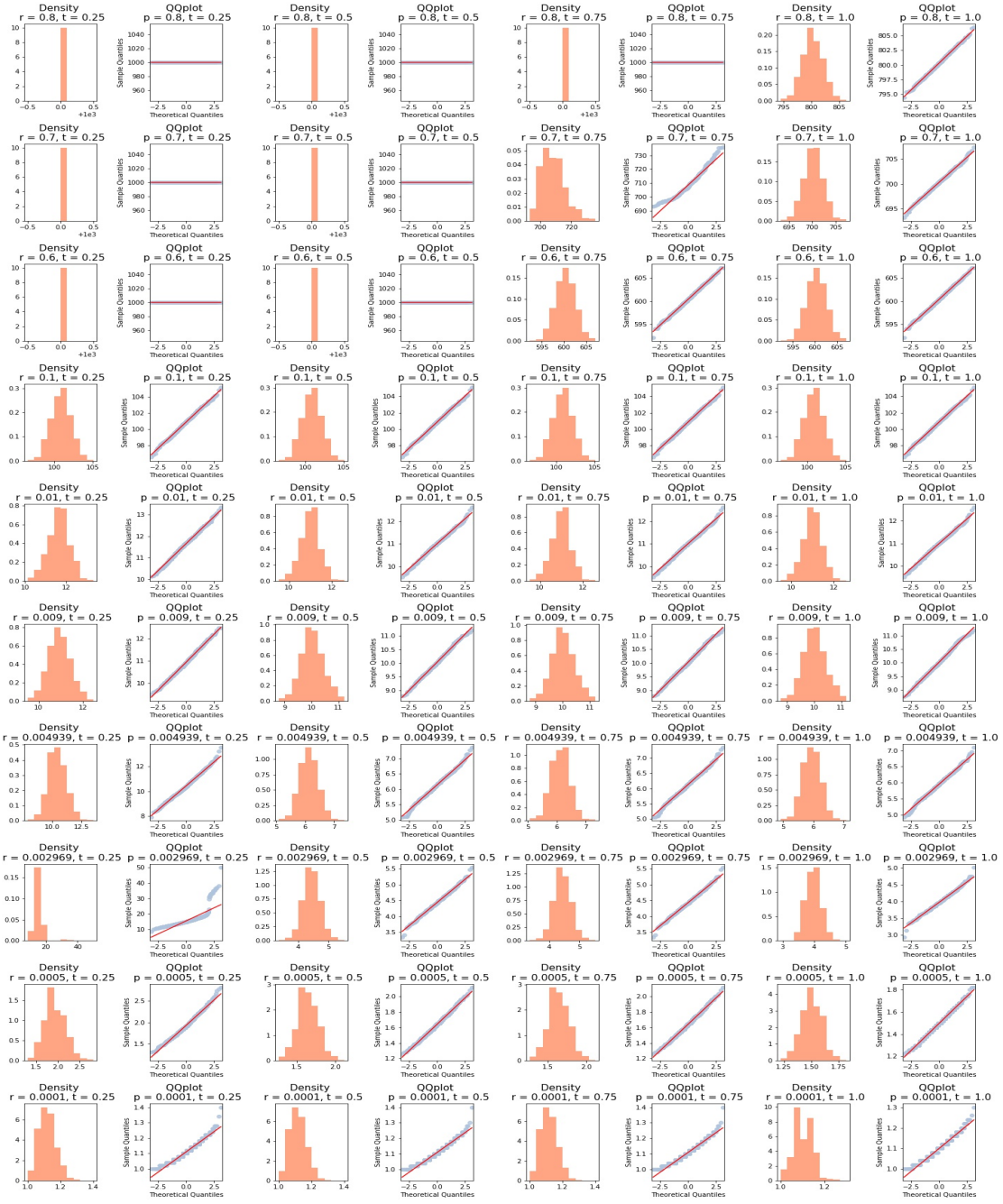


Figure A.20: $n = 1000$, undirected.

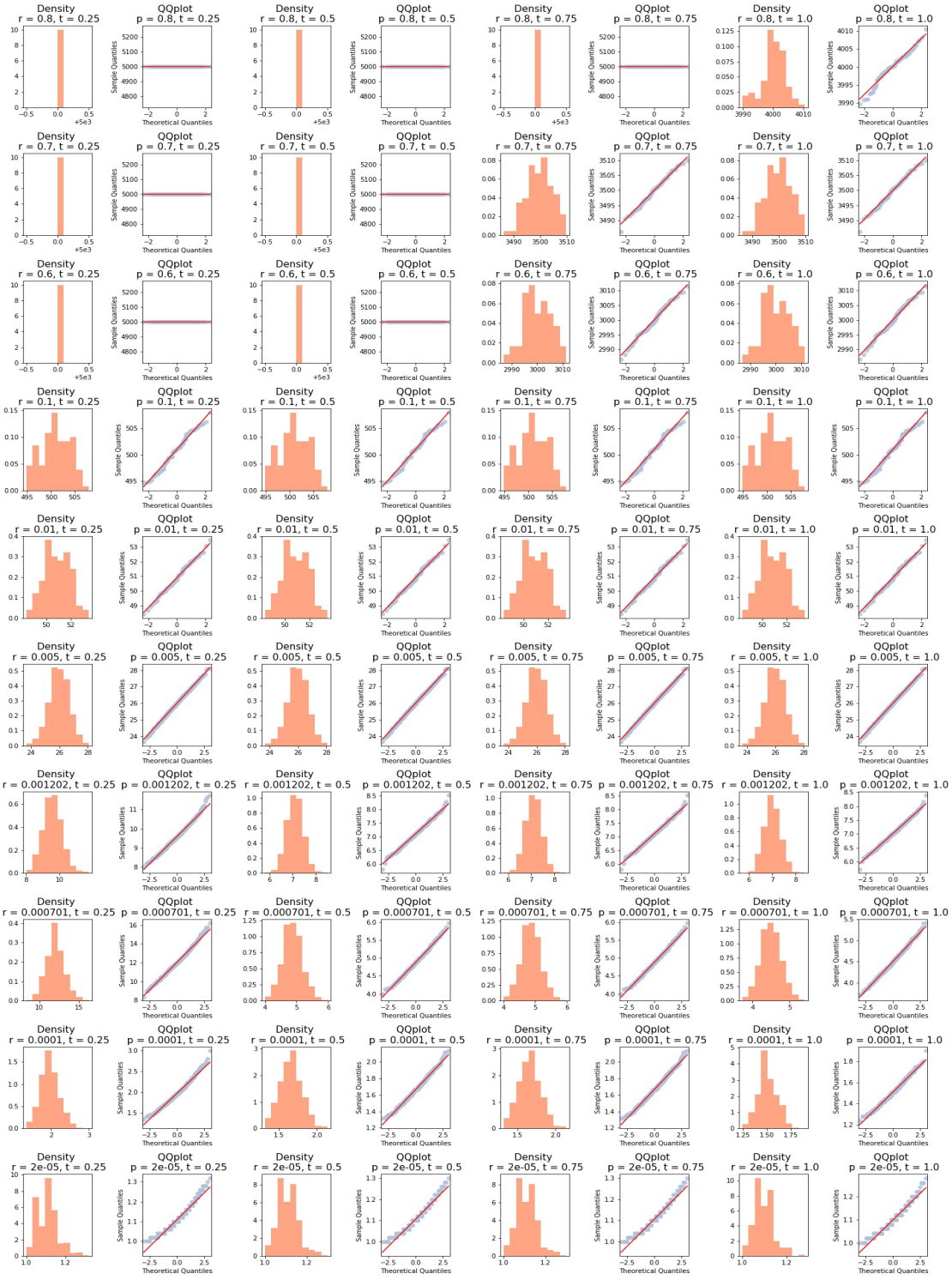


Figure A.21: $n = 5000$, undirected.

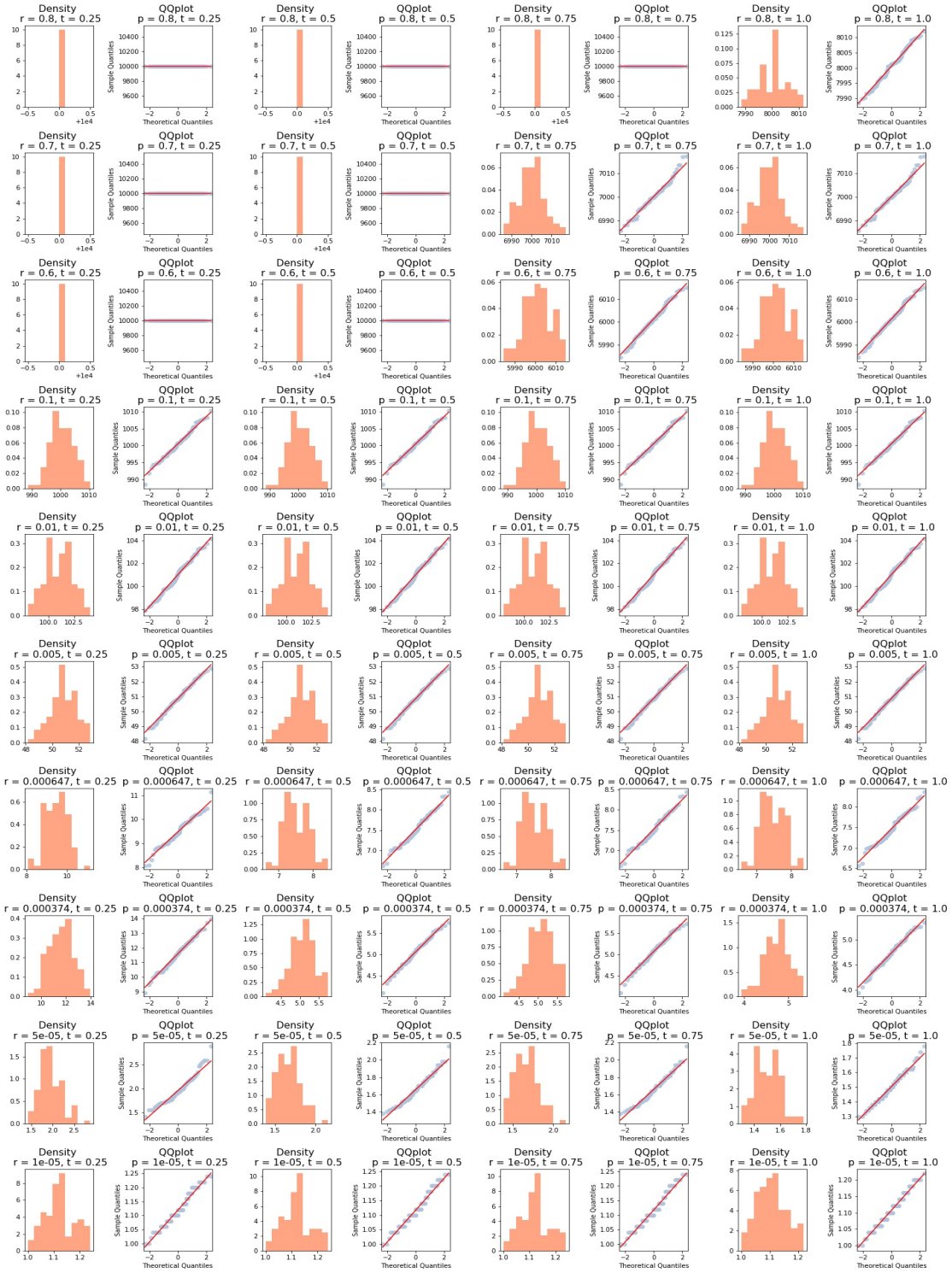


Figure A.22: $n = 10000$, undirected.

A.3.2 Refinement on high probabilities: inflection point.

Here represented the results for the refined t parameter (according to Second experiment phase 3, Section 4.3) and sample size 100.

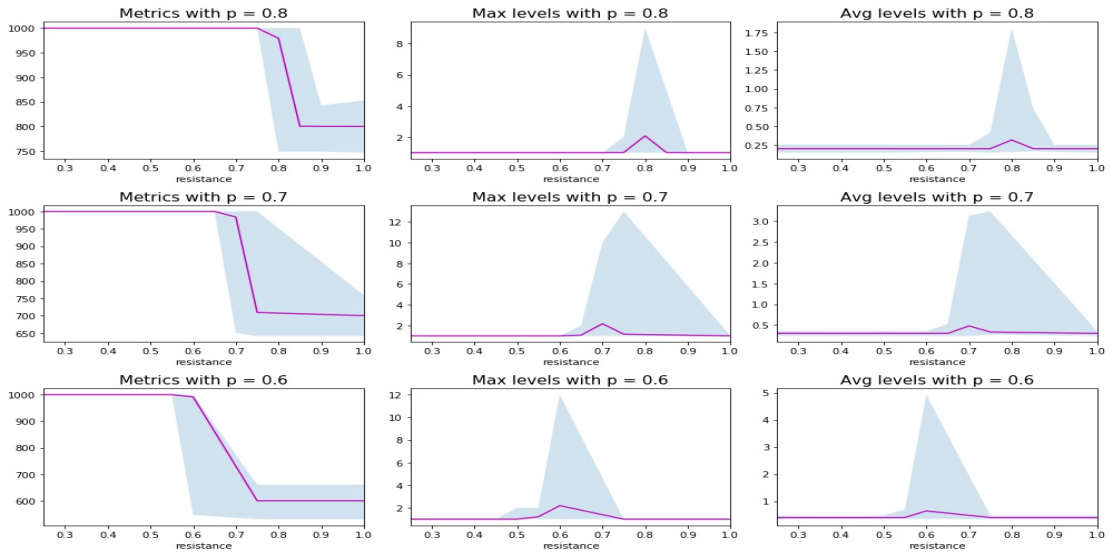


Figure A.23: $n = 1000$, undirected. Violet line: mean value, Blue area: [min, max] range.

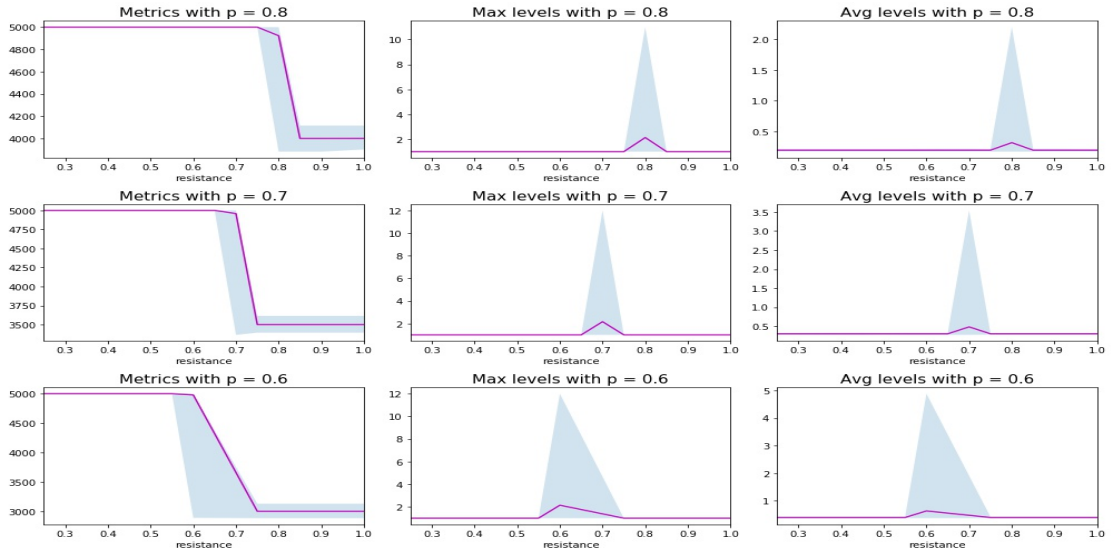


Figure A.24: $n = 5000$, undirected. Violet line: mean value, Blue area: [min, max] range.

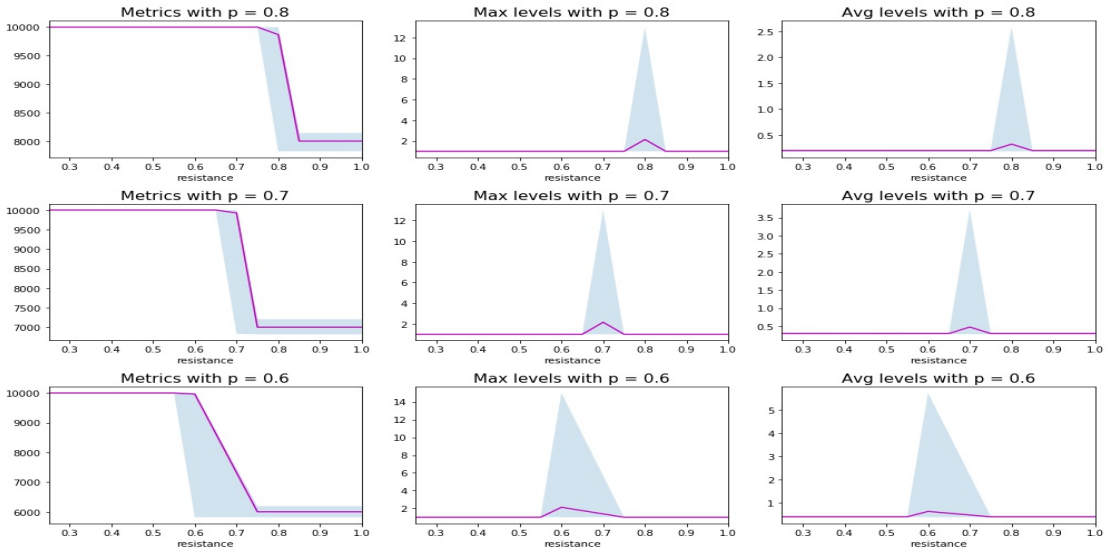


Figure A.25: $n = 10000$, undirected. Violet line: mean value, Blue area: [min, max] range.

A representation of the behaviour of the `metric` versus the `max_level` parameters is here provided to better characterize the inflection point.

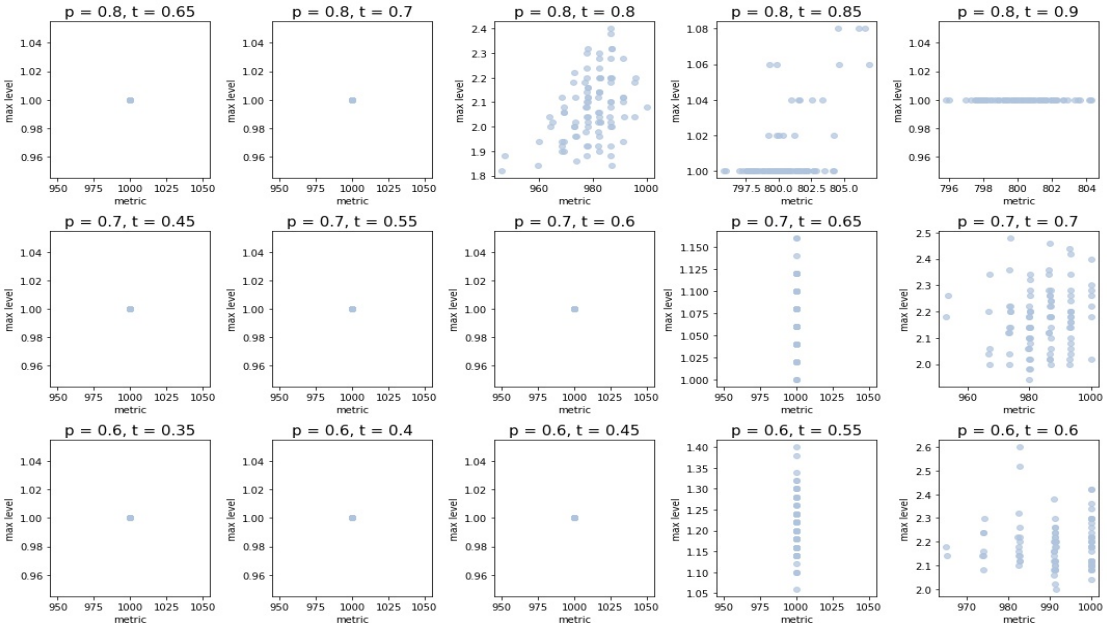


Figure A.26: $n = 1000$, undirected.

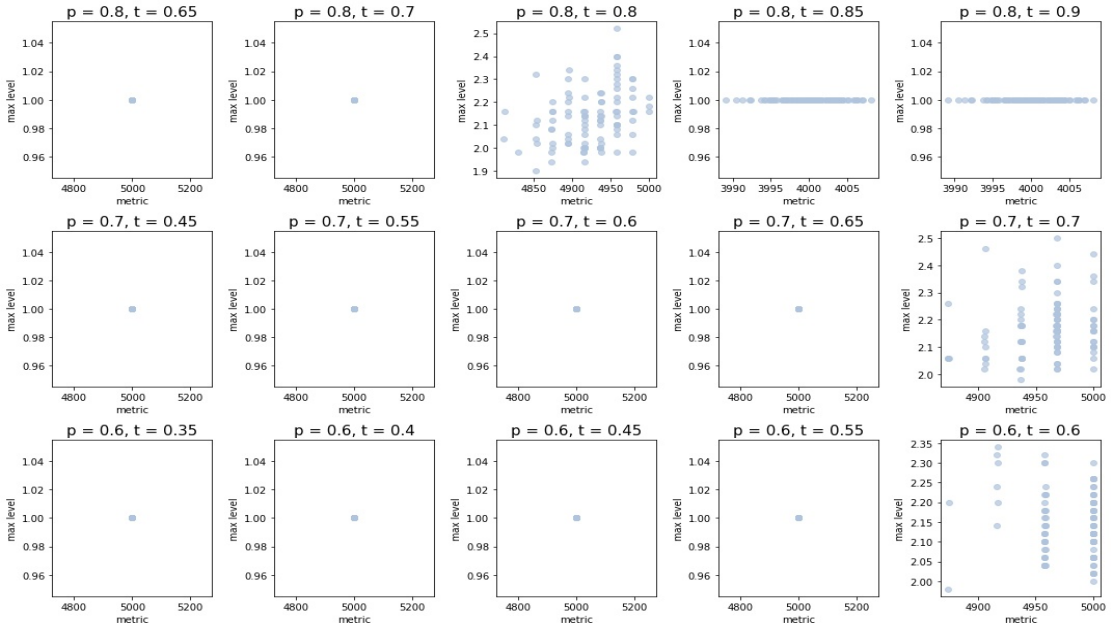


Figure A.27: $n = 5000$, undirected.

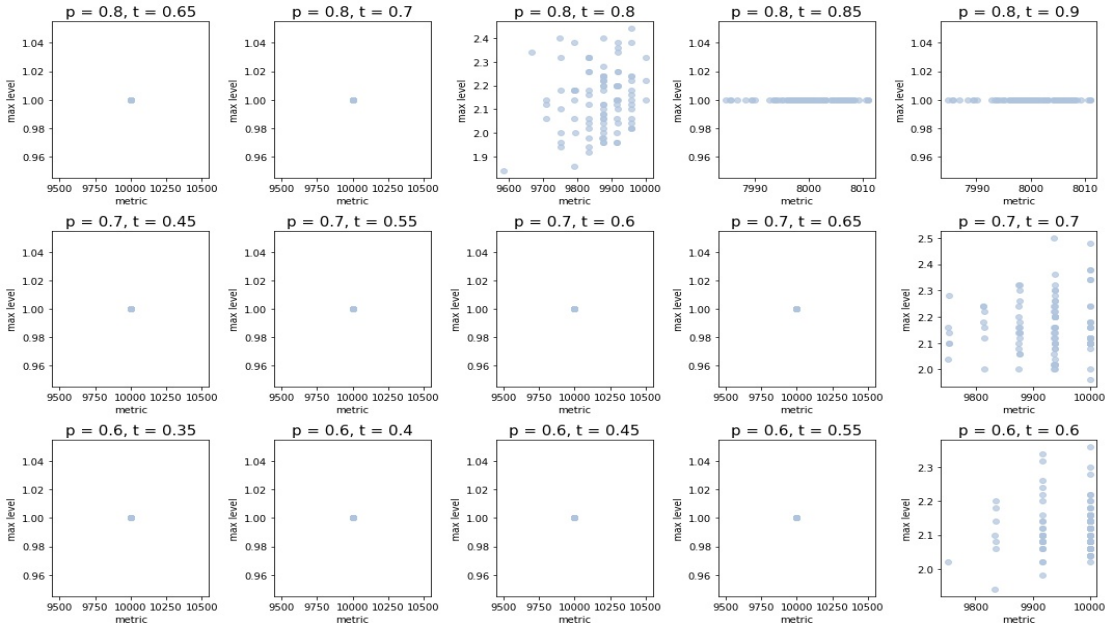
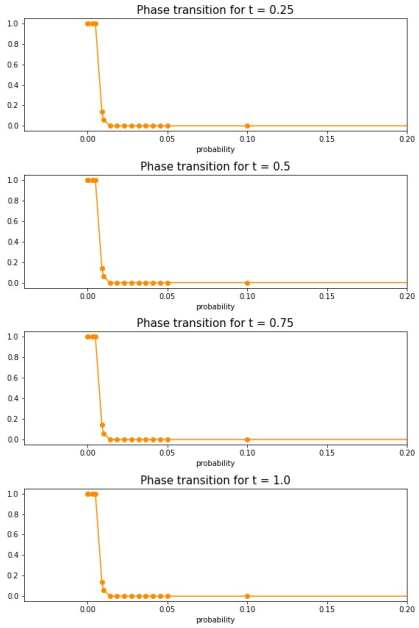


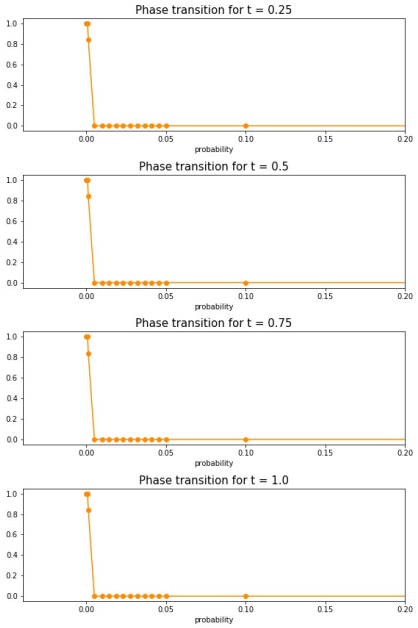
Figure A.28: $n = 10000$, undirected.

A.4 Phase transitions with neighbour threshold

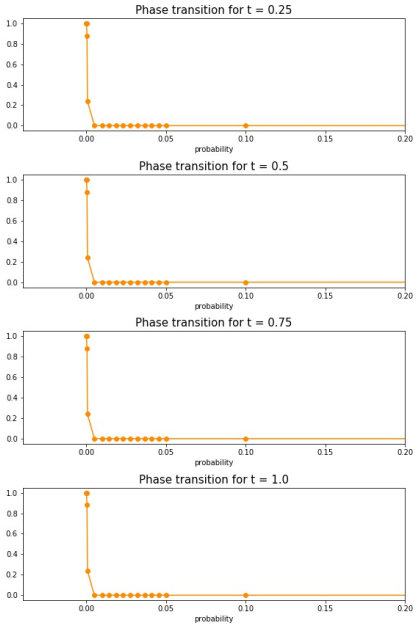
A.4.1 $\mathcal{P}_0$ phase transitions with neighbour threshold



(a) $n = 1000$, undirected.



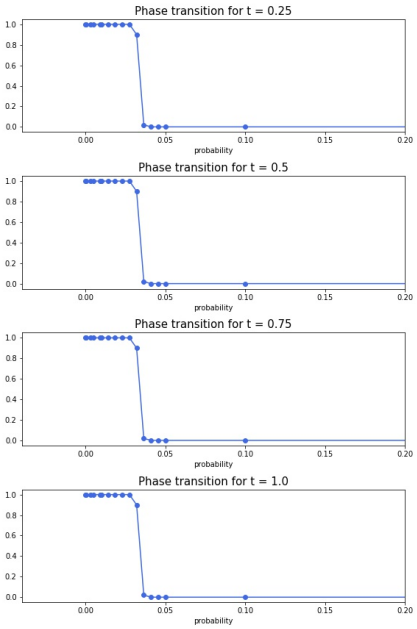
(b) $n = 5000$, undirected.



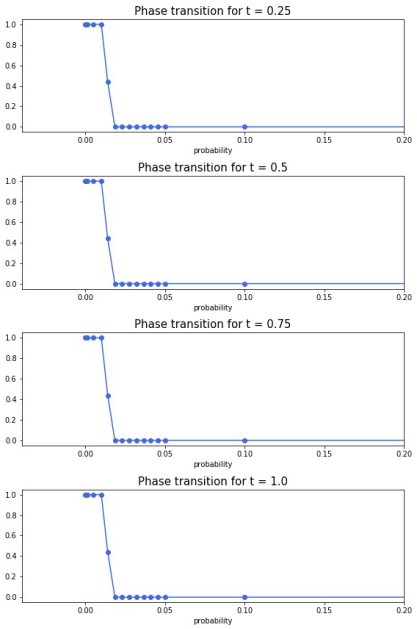
(c) $n = 10000$, undirected.

Figure A.29: Representation of the mean truth value assumed by the $\mathcal{P}_0$ property, mean computed on all the realizations.

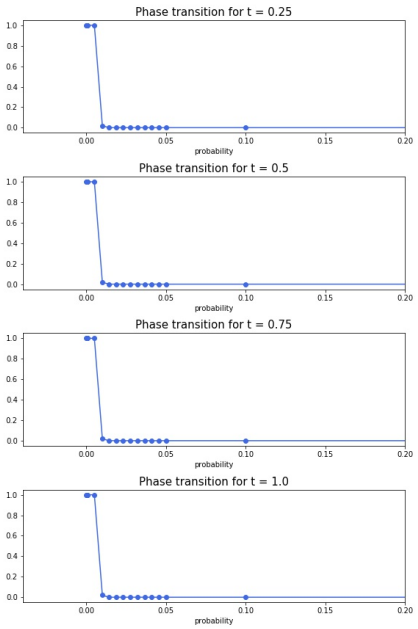
A.4.2 $\mathcal{P}_1$ phase transitions with neighbour threshold



(a) $n = 1000$, undirected.



(b) $n = 5000$, undirected.



(c) $n = 10000$, undirected.

Figure A.30: Representation of the mean truth value assumed by the $\mathcal{P}_1$ property, mean computed on all the realizations.

Appendix B

Images for Random Geometric Graphs

B.1 LTR with max neighbour threshold

Outputs per node computed as the average on all the realizations, here only shown for the initial probability and t sets.

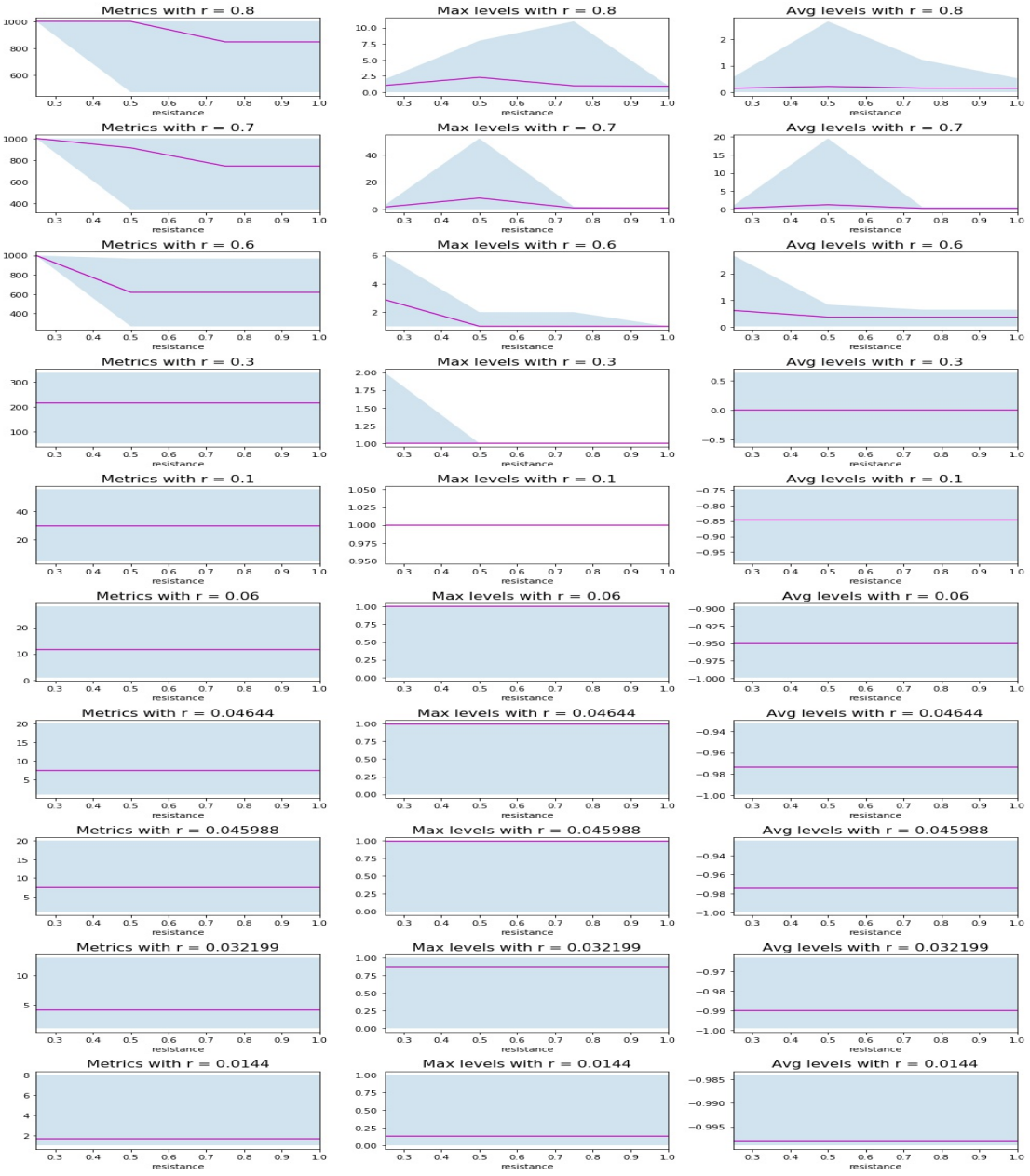


Figure B.1: $n = 1000$. Violet line: mean value, Blue area: [min, max] range.

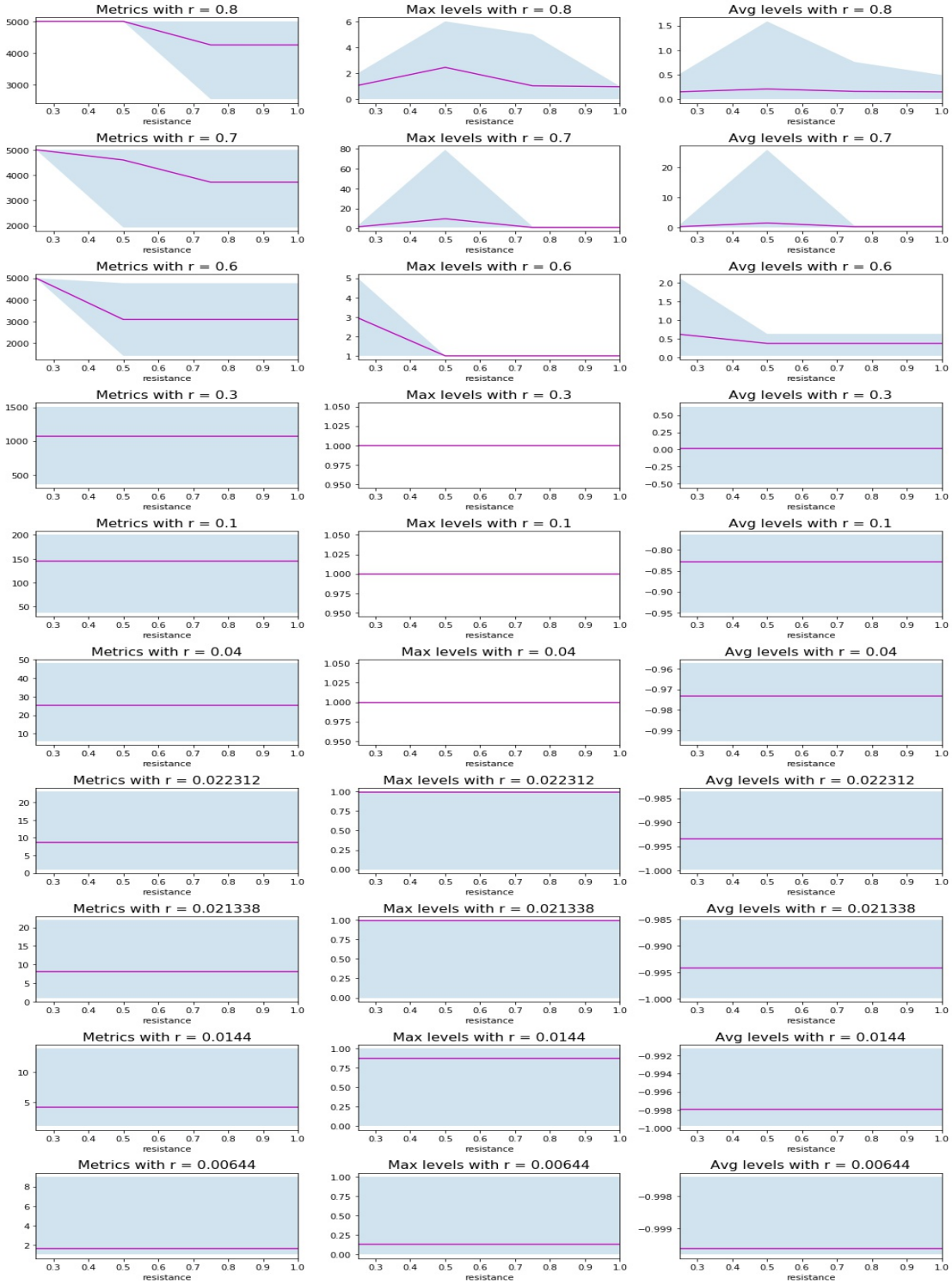


Figure B.2: $n = 5000$. Violet line: mean value, Blue area: [min, max] range.

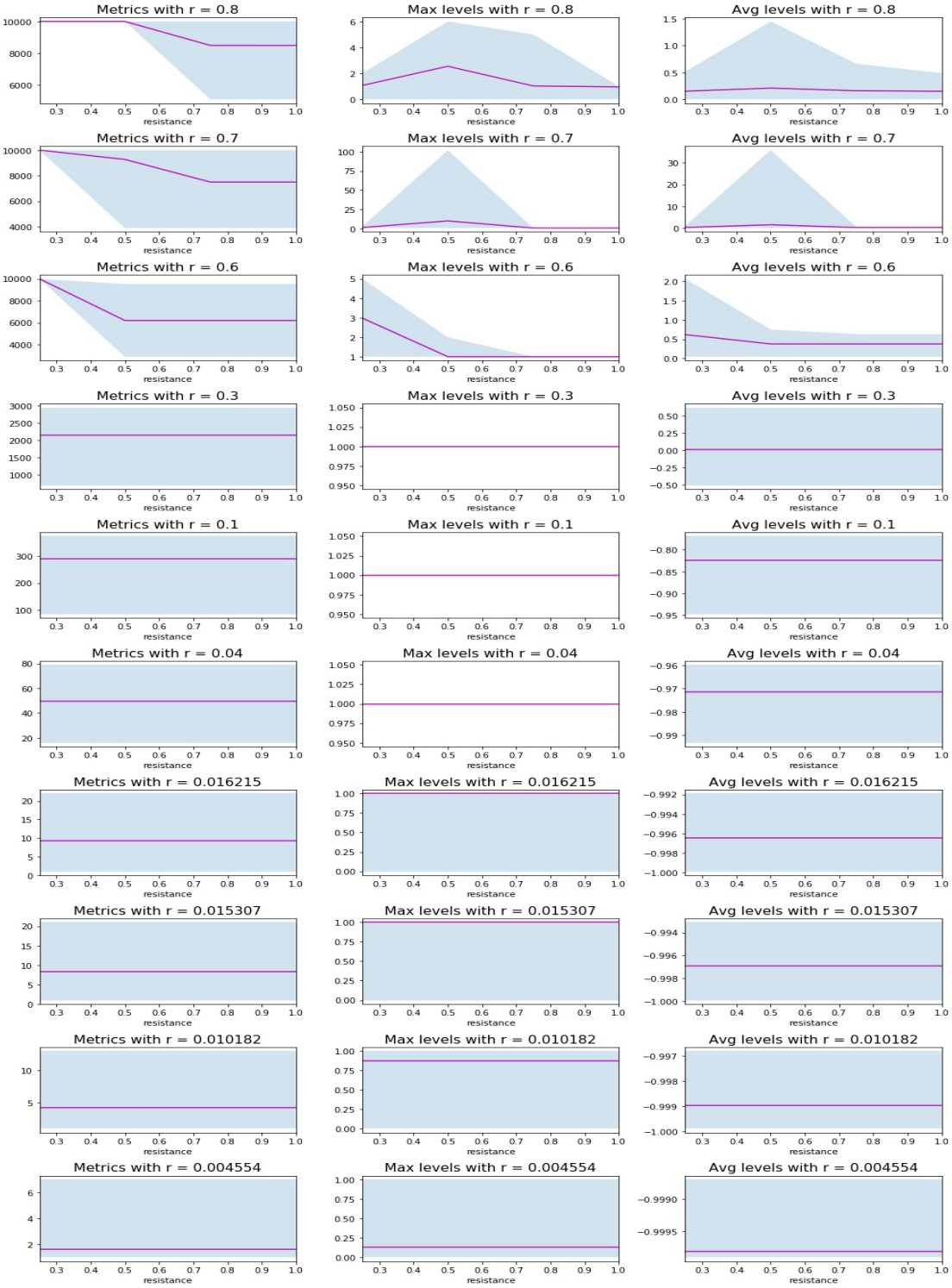


Figure B.3: $n = 10000$. Violet line: mean value, Blue area: [min, max] range.

B.1.1 Distribution the LTR with max neighbour threshold

Here the approximated density of the `metric` parameter is given, plus a normality test (QQplot). x axis: $\mathbb{E}_{i,k}$(`metric`); y axis: probability.

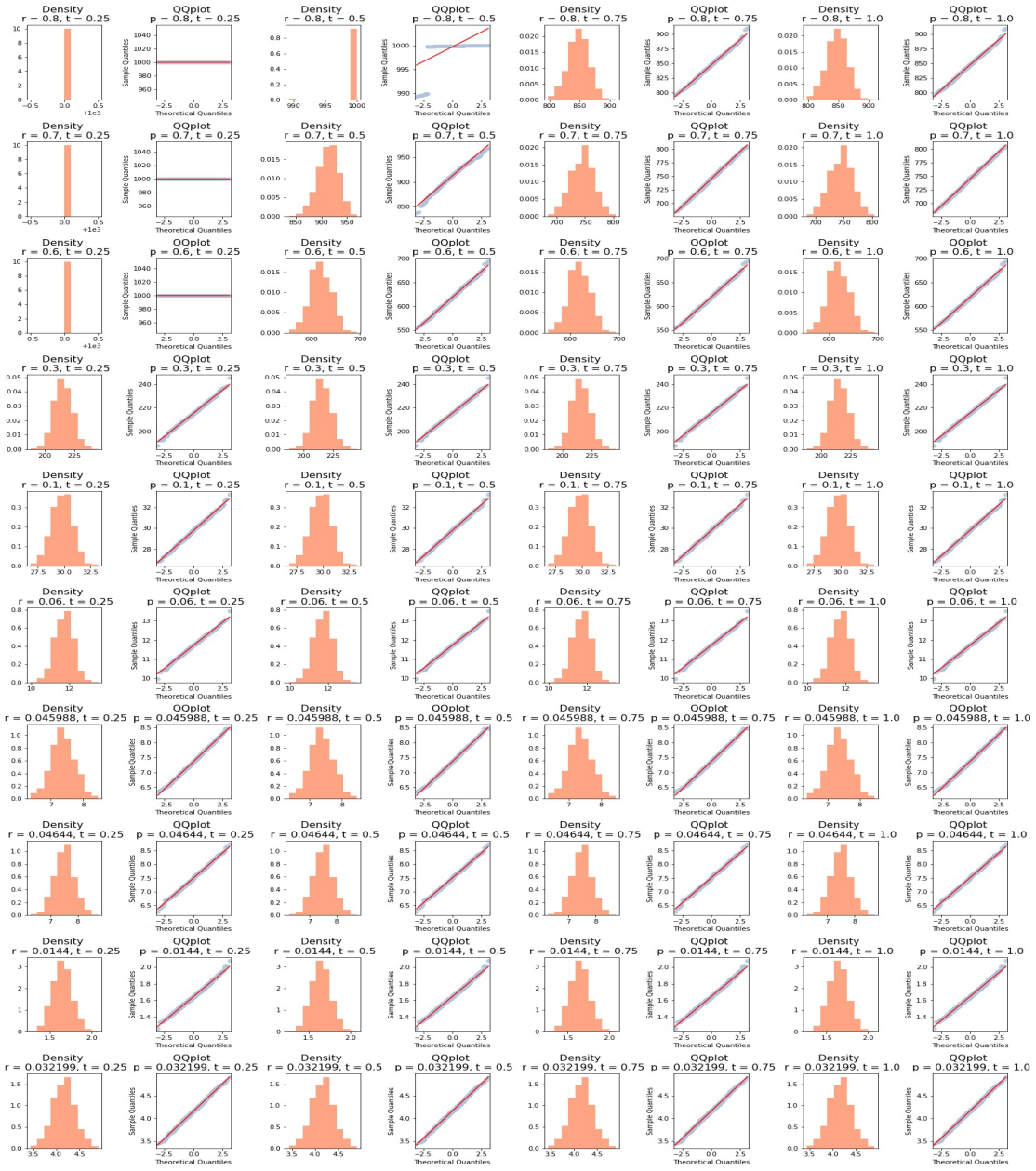


Figure B.4: $n = 1000$.

Figure B.5: $n = 5000$.

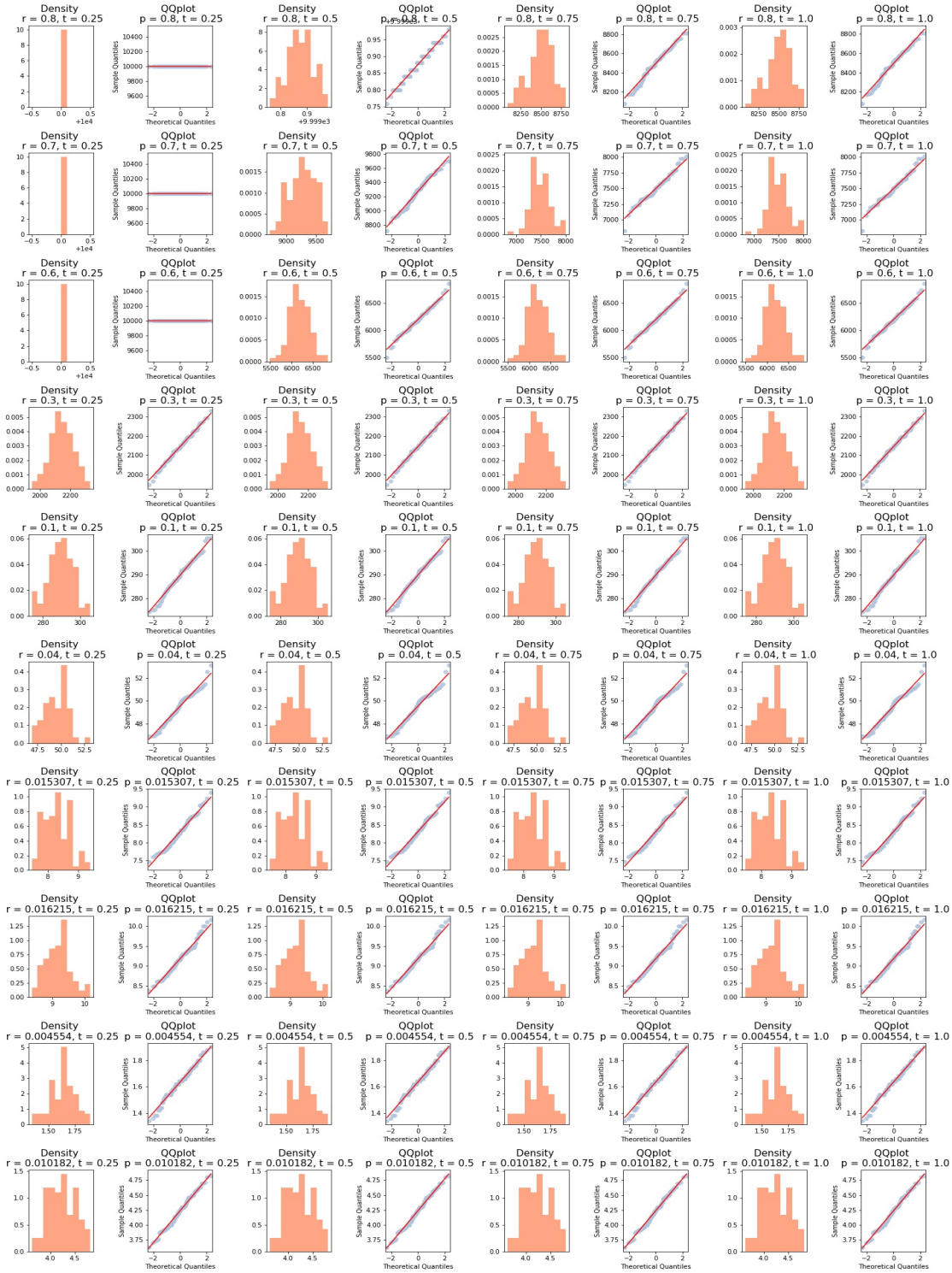


Figure B.6: $n = 10000$.

B.1.2 Inflection point: Metric vs Maxlevel

A representation of the behaviour of the `metric` versus the `max_level` parameters is here provided to better characterize the inflection point.

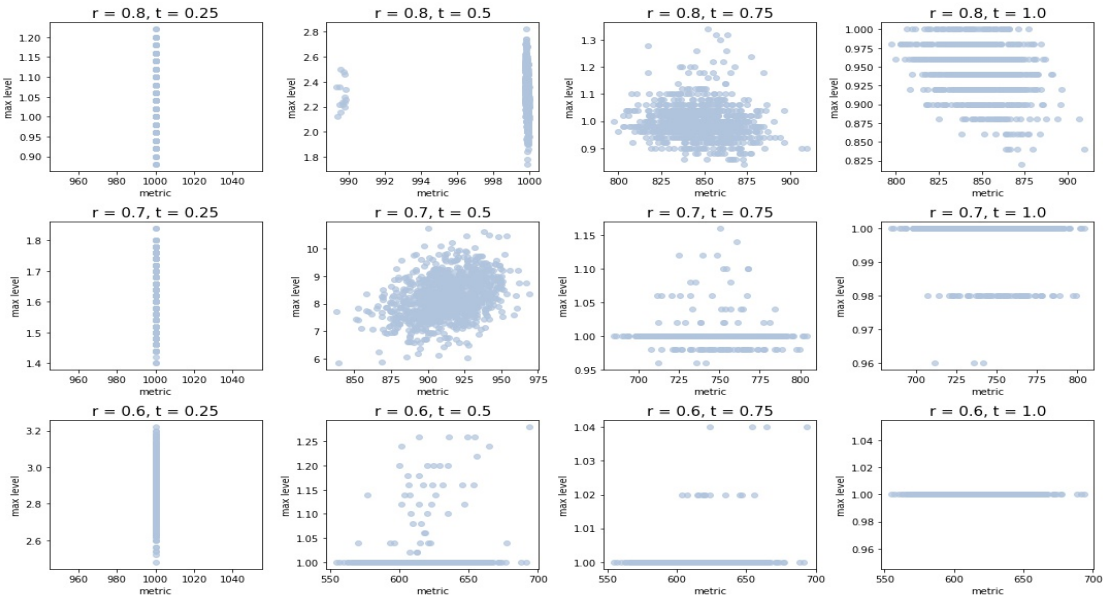


Figure B.7: $n = 1000$.

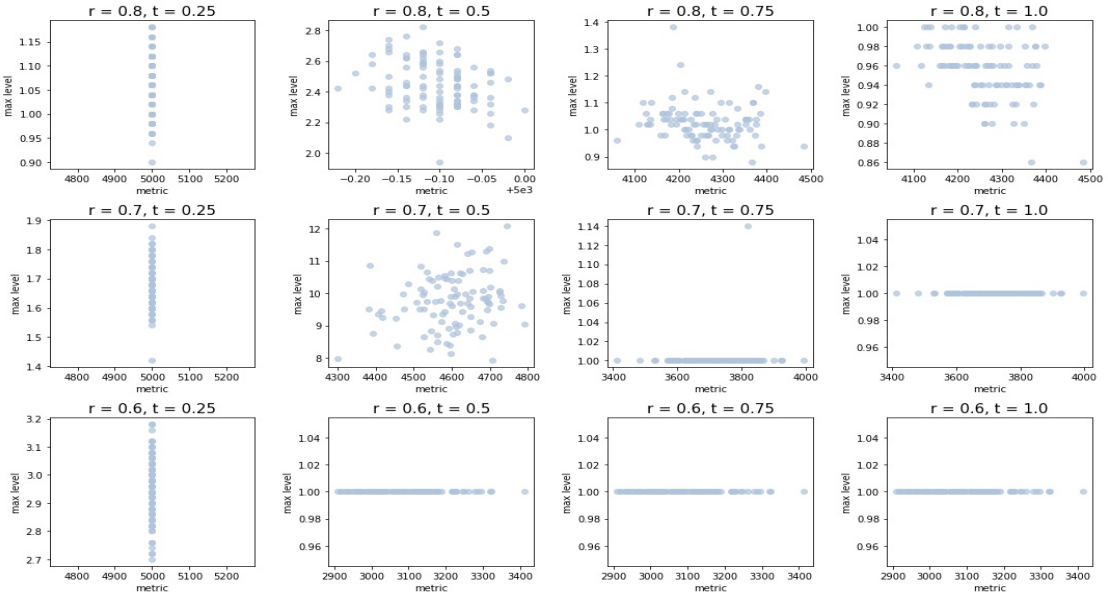


Figure B.8: $n = 5000$.

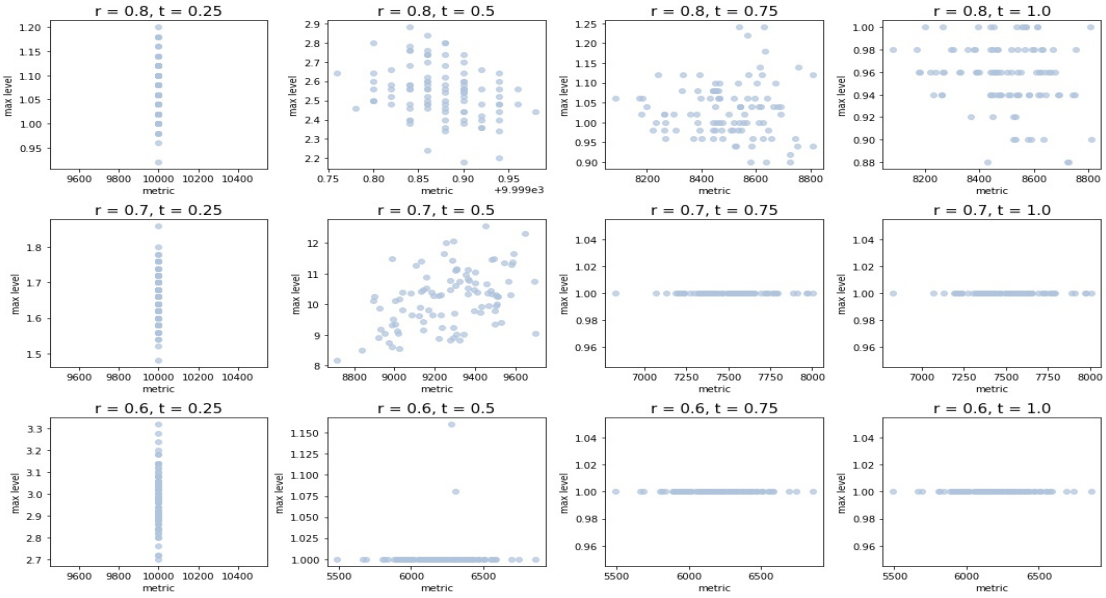
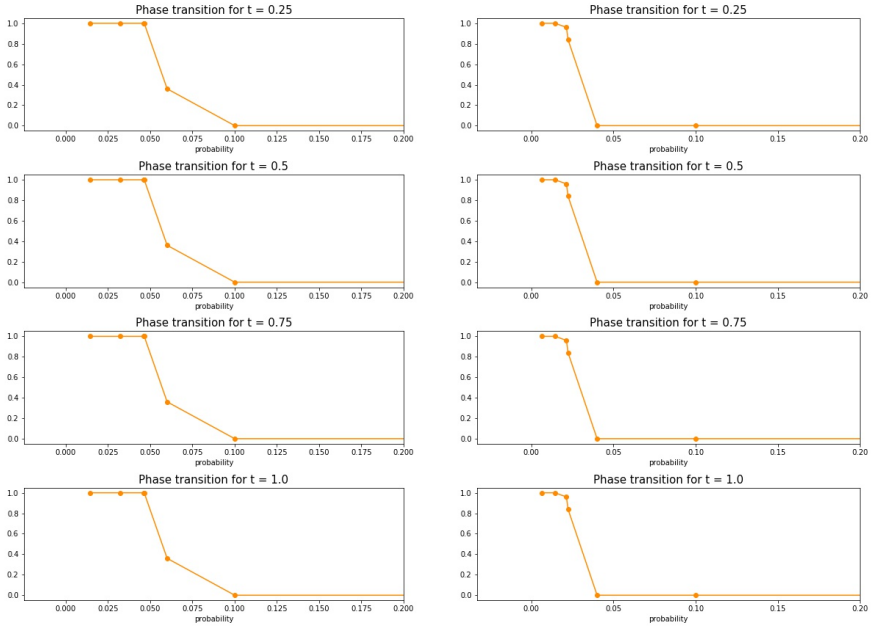


Figure B.9: $n = 10000$.

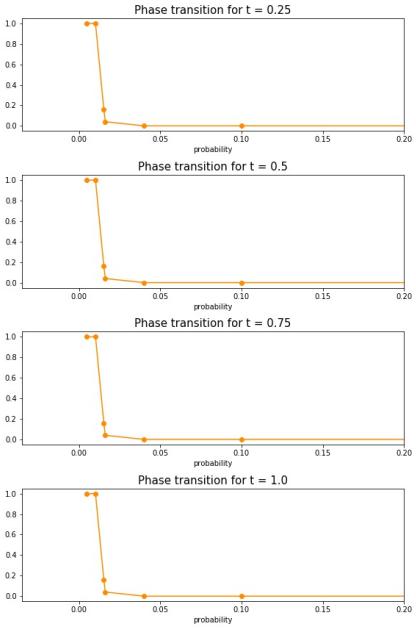
B.2 Phase transitions with max neighbour threshold

B.2.1 $\mathcal{P}_0$ phase transitions with max neighbour threshold



(a) $n = 1000$.

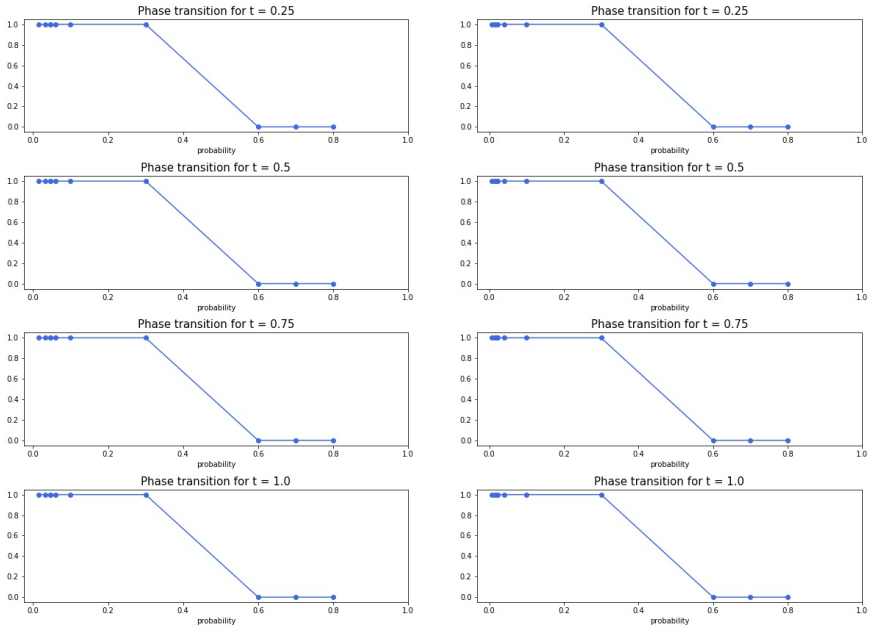
(b) $n = 5000$.



(c) $n = 10000$.

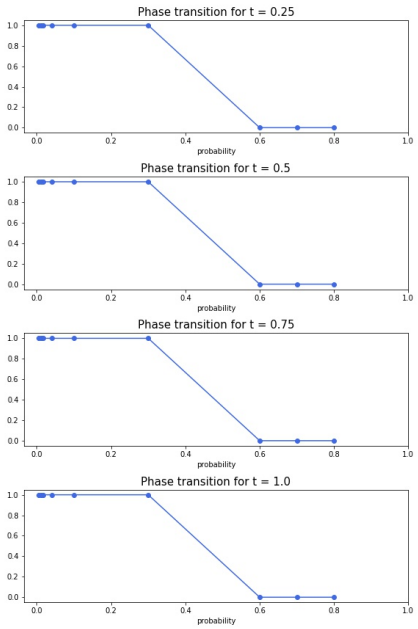
Figure B.10: Representation of the mean truth value assumed by the $\mathcal{P}_0$ property, mean computed on all the realizations.

B.2.2 $\mathcal{P}_1$ phase transitions with max neighbour threshold



(a) $n = 1000$.

(b) $n = 5000$.



(c) $n = 10000$.

Figure B.11: Representation of the mean truth value assumed by the $\mathcal{P}_1$ property, mean computed on all the realizations.

B.3 LTR with neighbour threshold

Outputs per node computed as the average on all the realizations, here only shown for the initial probability and t sets.

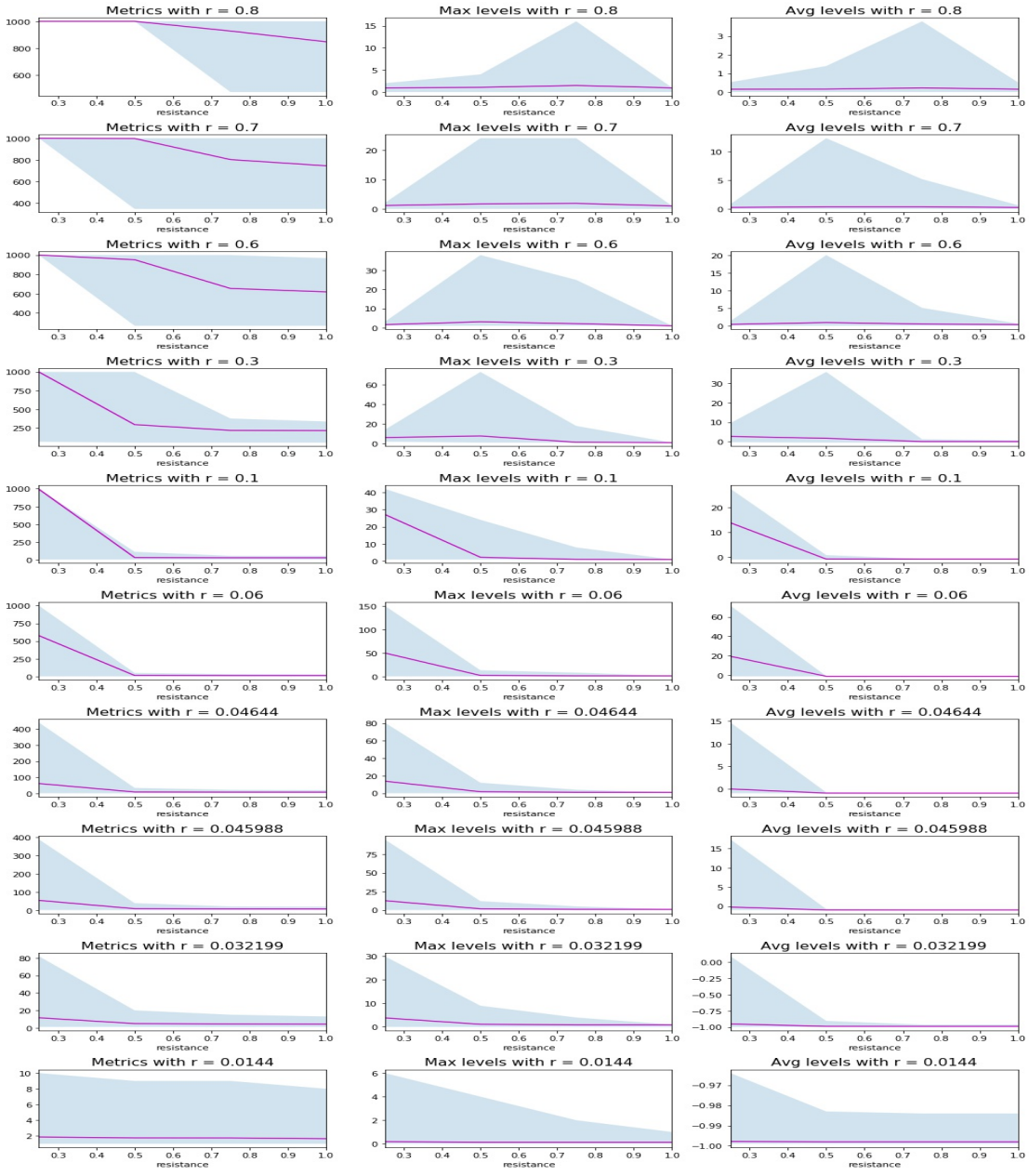


Figure B.12: $n = 1000$. Violet line: mean value, Blue area: [min, max] range.

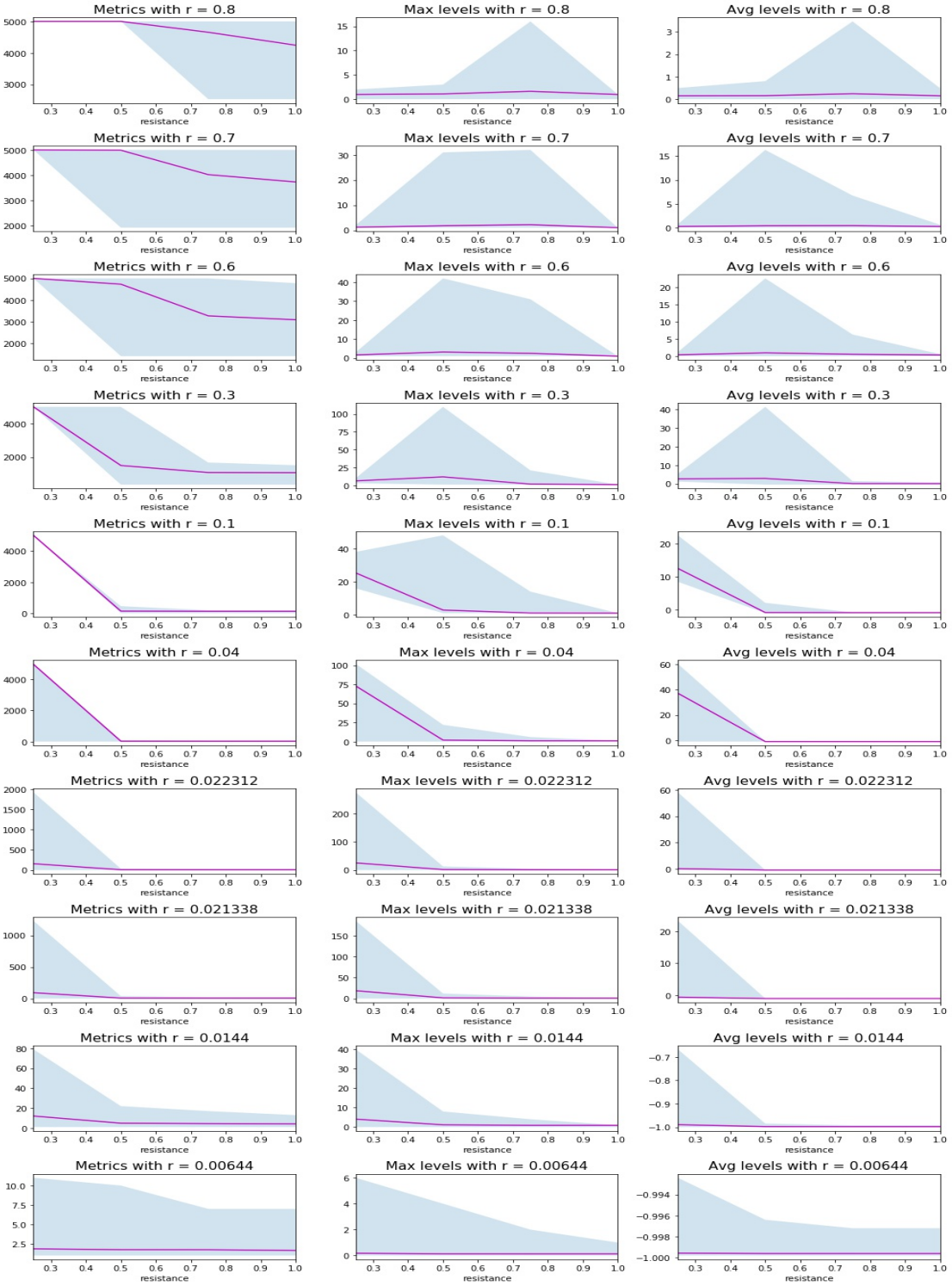


Figure B.13: $n = 5000$. Violet line: mean value, Blue area: [min, max] range.

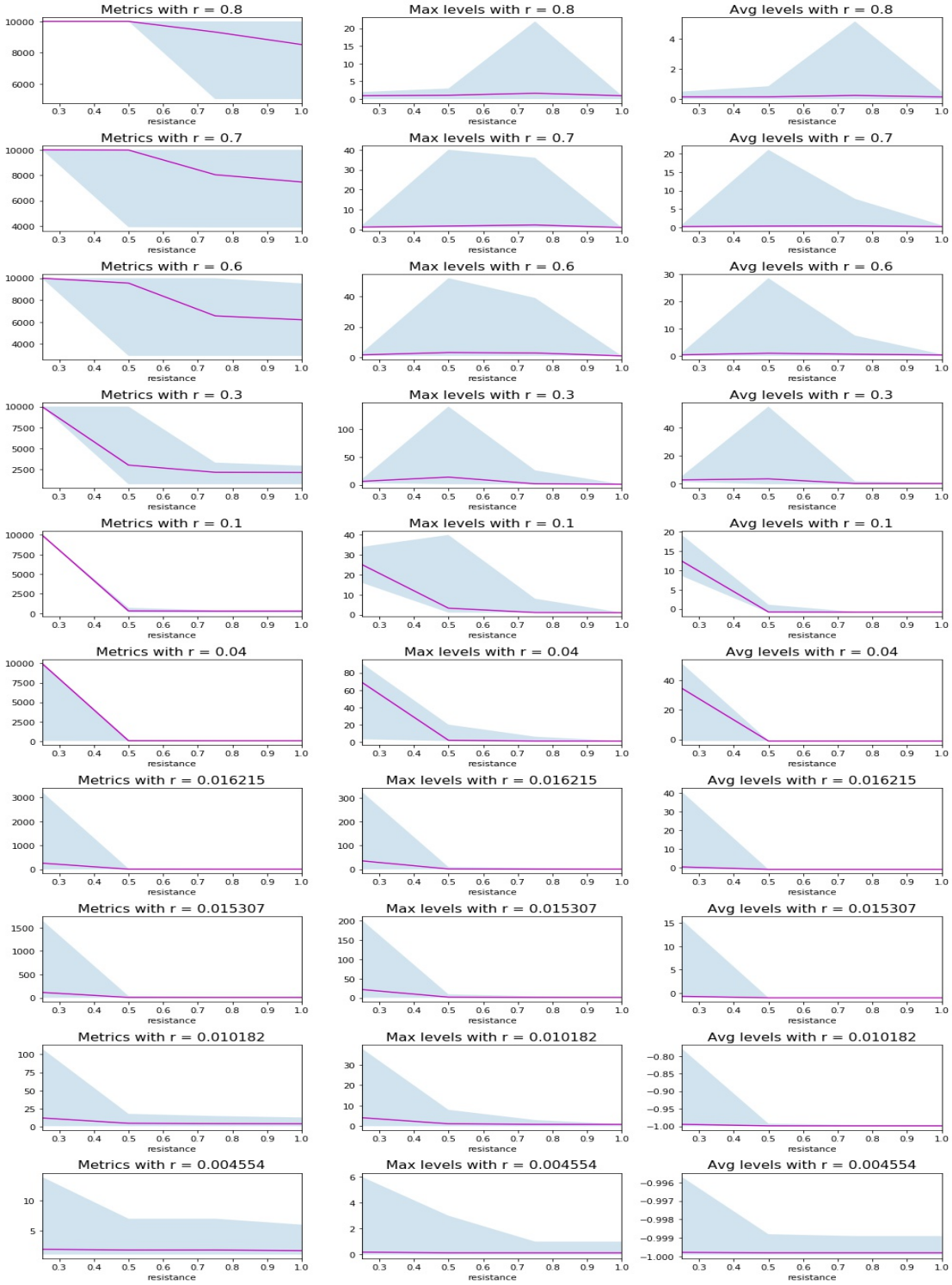


Figure B.14: $n = 10000$. Violet line: mean value, Blue area: [min, max] range.

B.3.1 Distribution the LTR with neighbour threshold

Here the approximated density of the `metric` parameter is given, plus a normality test (QQplot). x axis: $\mathbb{E}_{i,k}$(`metric`); y axis: probability.

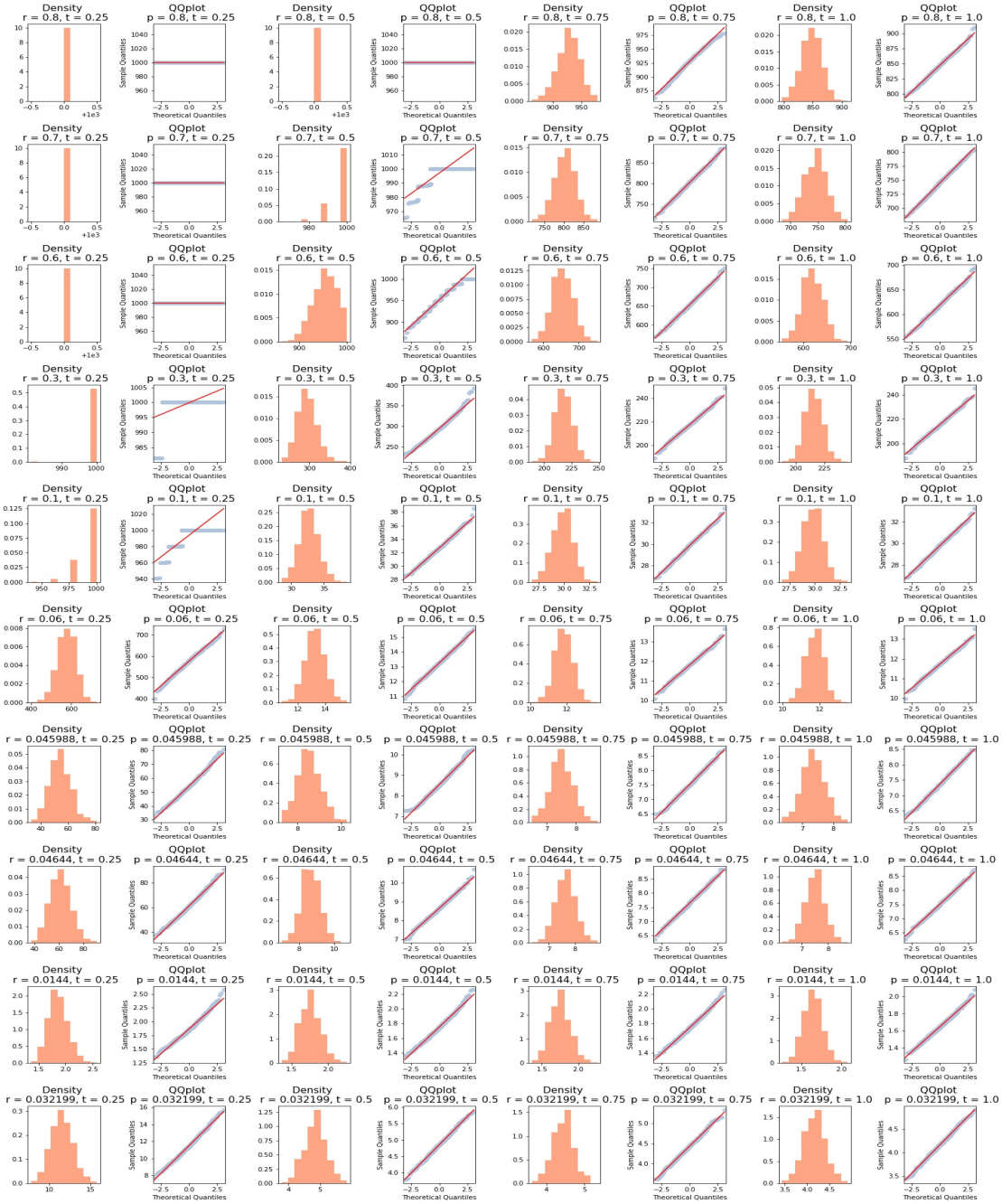


Figure B.15: $n = 1000$.

Figure B.16: $n = 5000$.

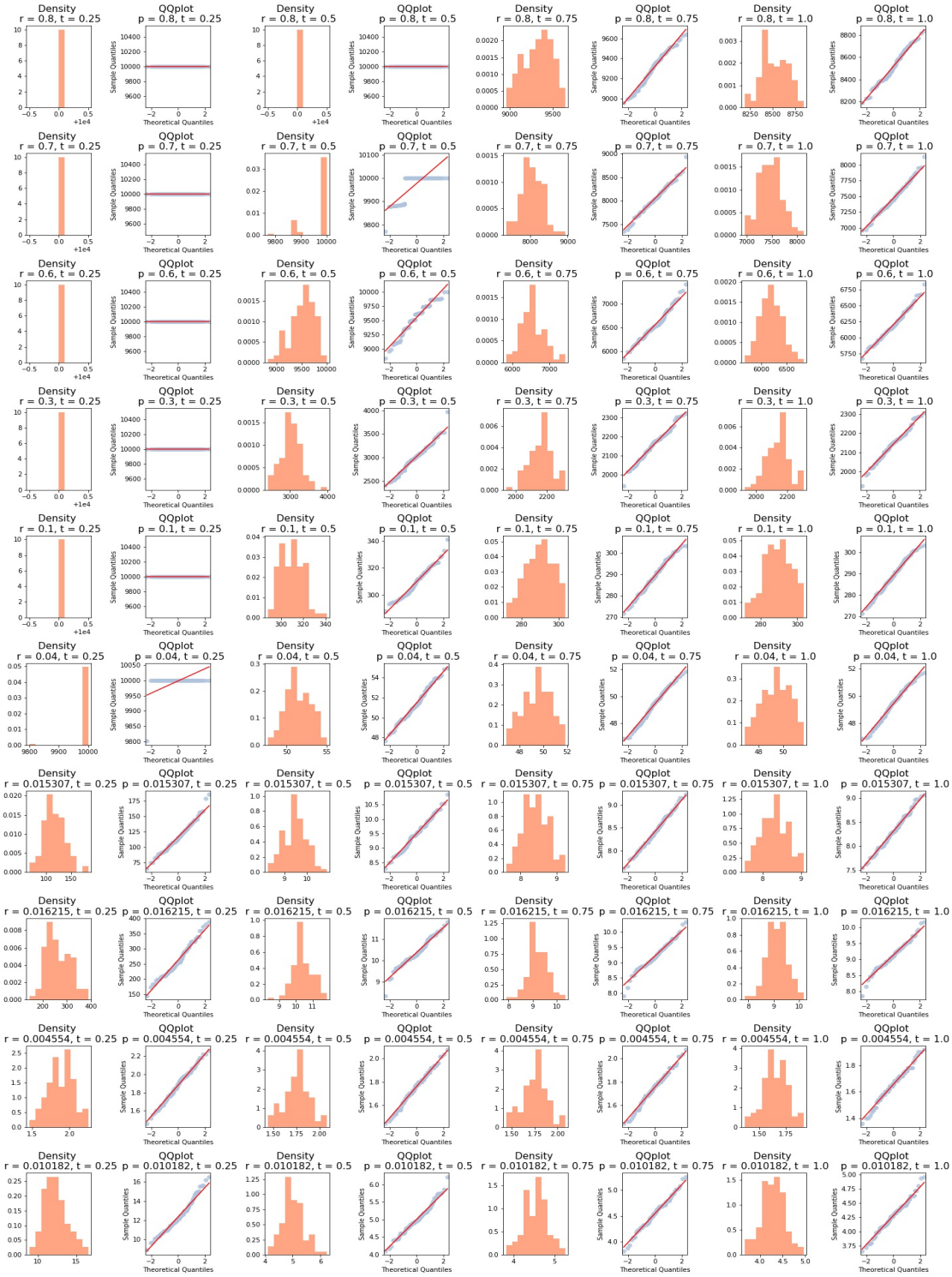


Figure B.17: $n = 10000$.

B.3.2 Inflection point: Metric vs Maxlevel

A representation of the behaviour of the `metric` versus the `max_level` parameters is here provided to better characterize the inflection point.

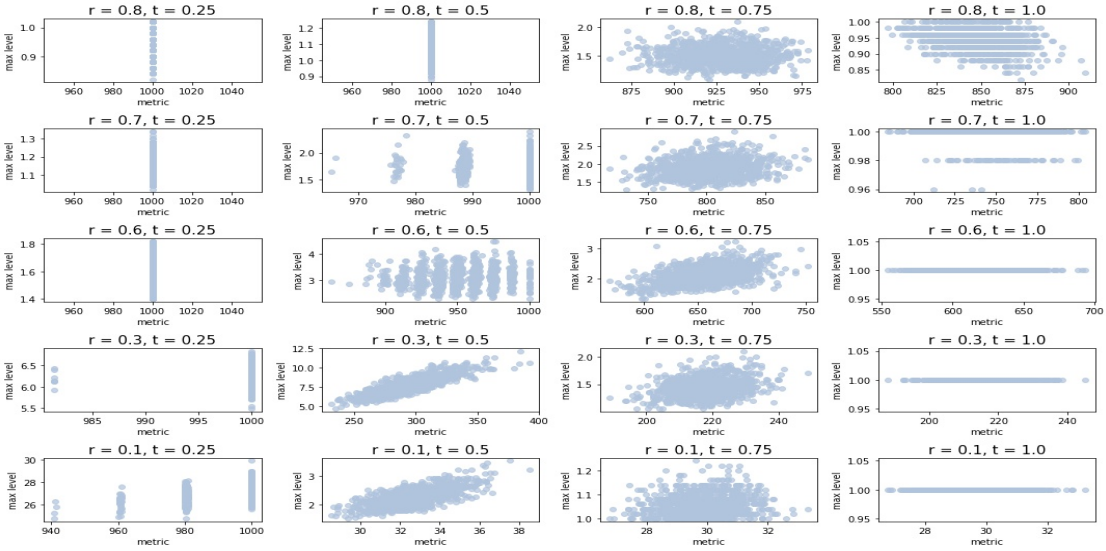


Figure B.18: $n = 1000$.

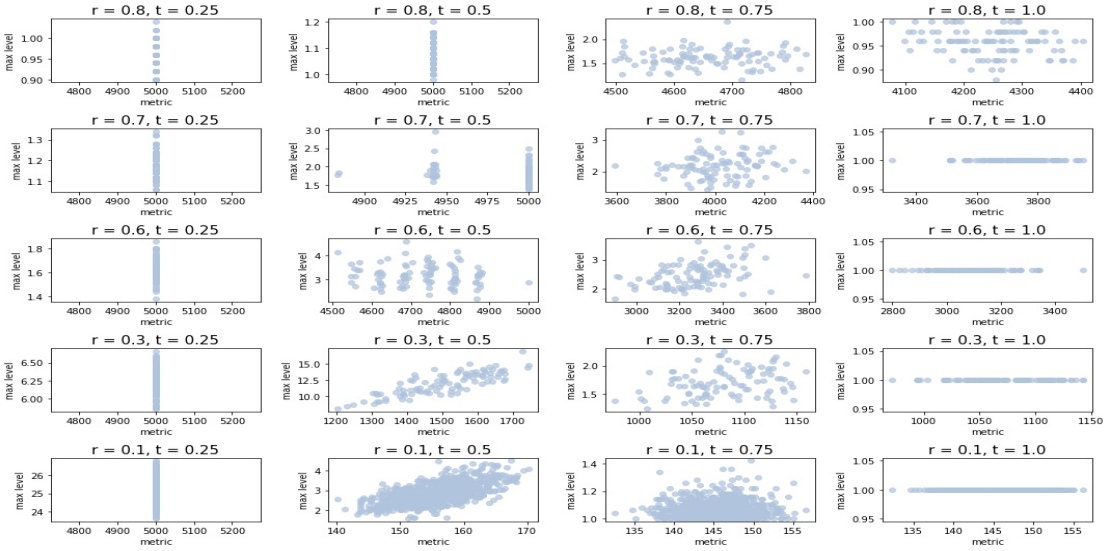


Figure B.19: $n = 5000$.

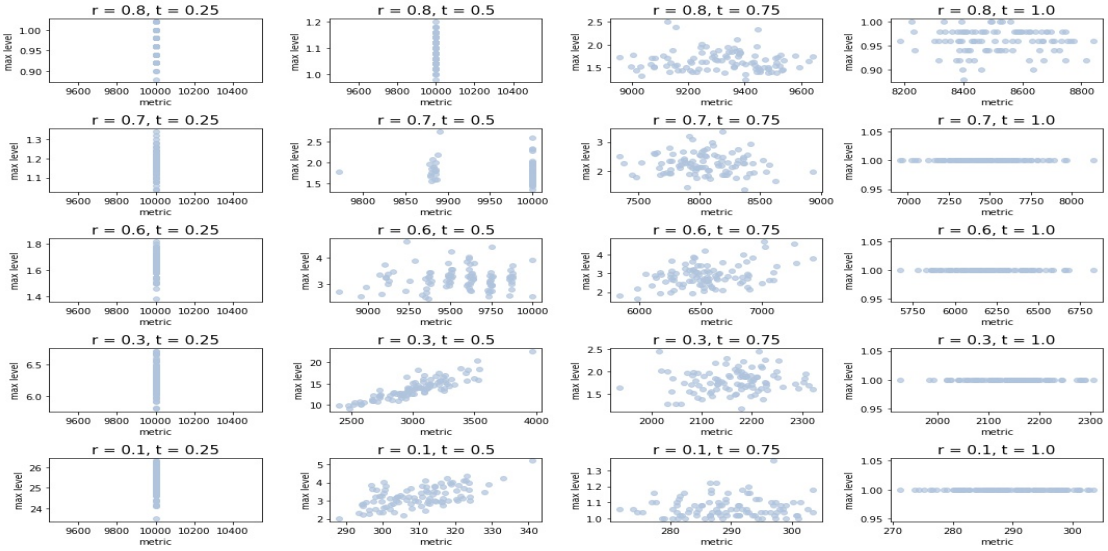
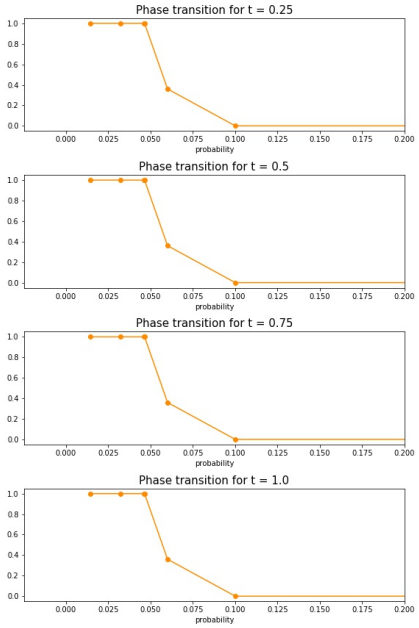


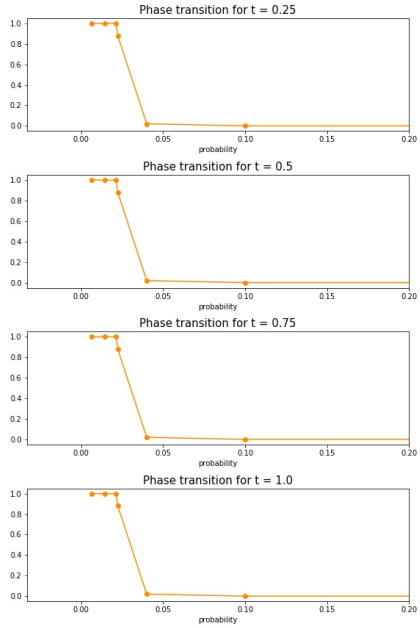
Figure B.20: $n = 10000$.

B.4 Phase transitions with neighbour threshold

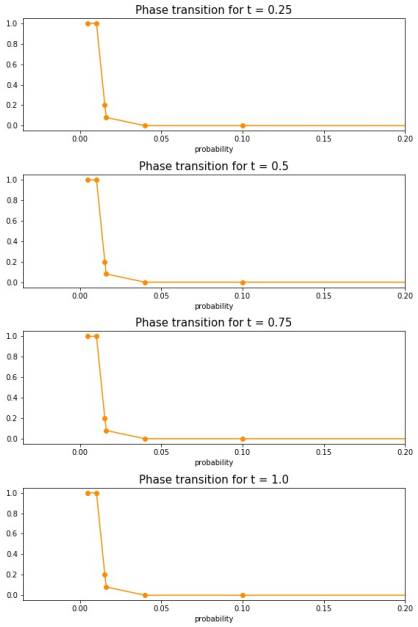
B.4.1 $\mathcal{P}_0$ phase transitions with neighbour threshold



(a) $n = 1000$.



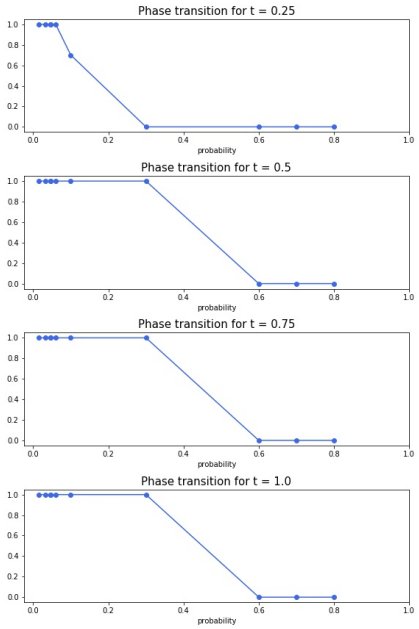
(b) $n = 5000$.



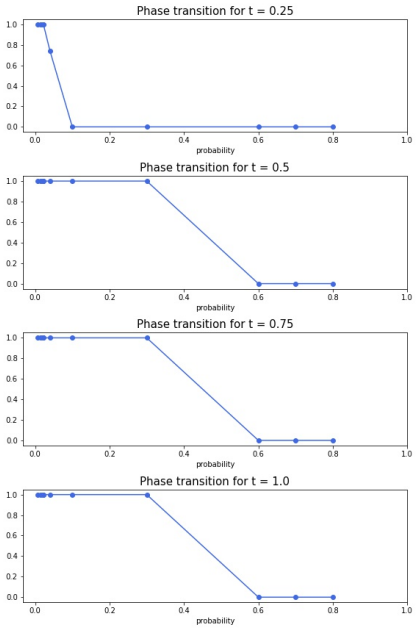
(c) $n = 10000$.

Figure B.21: Representation of the mean truth value assumed by the $\mathcal{P}_0$ property, mean computed on all the realizations.

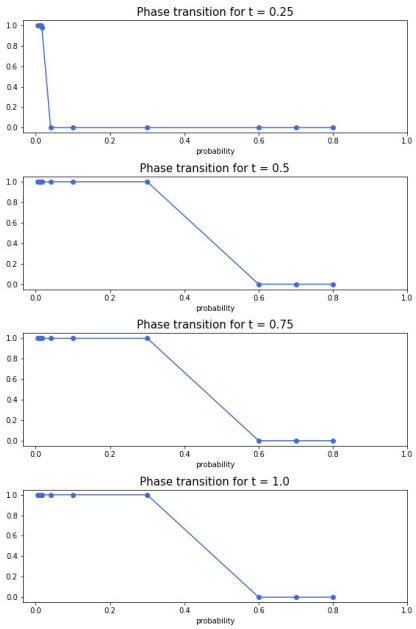
B.4.2 $\mathcal{P}_1$ phase transitions with neighbour threshold



(a) $n = 1000$.



(b) $n = 5000$.



(c) $n = 10000$.

Figure B.22: Representation of the mean truth value assumed by the $\mathcal{P}_1$ property, mean computed on all the realizations.