Peaks and valleys in the size distribution of shortest path subgraphs

Sergey Kirgizov and Clémence Magnien

firstname.lastname@lip6.fr
UPMC, LIP6, UMR 7606, 75252 Paris, France
CNRS, UMR 7606, 75252 Paris, France

June 27, 2014

Abstract

Considering a random graph $G(n, p)$ we denote by $SPS(u, v)$ the subgraph of all shortest paths between two vertices u and v. We show that the size of $SPS(u, v)$ follows a nontrivial probability law with several local maximum values. In the random graphs with constant density p the number of vertices in SPS is equal to 2 (with the probability p) or concentrated around np^2 (with the probability $1 - p$) as n goes to infinity. In sparse random graphs with unbounded mean degree ($p \to 0, n \to \infty, np \to \infty$) we have similar two-peak distribution. Also we approximate the expected number of vertices in $SPS(u, v)$, when the distance between u and v is known (we give the exact distribution, when the distance is equal to 2).

1 Background and introduction

Random graphs attract great attention of researchers from various fields, such graphs are the basis for many models of real-world networks. Our work is motivated by the analysis of the Internet topology measurements for which a model has been proposed [Magnien et al. 2013; Medem et al. 2012]. The basic operation for measuring the topology of the Internet consists in observing a single path between two vertices, with tools such as `traceroute` [Latapy et al. 2011]. In order to understand the hidden network structure between two computers in general network, we develop a theoretical approach using classical random graphs.

We consider a random graph $G(n, p)$, where n is the number of vertices and each edge is included in the graph with probability p independent from every other edge. This notion of a random graph goes back to the works [Rapoport 1948; Solomonoff and Rapoport 1951; Erdős and Rényi 1959], and [Gilbert 1959]. Massive numerical simulations was performed in order to estimate the average proportion of edges that lie on all shortest paths from a given vertex to all other vertices [Guillaume and Latapy 2005], which represents the fraction of links that it is possible to observe in the Internet using classical measurement tools. Guillaume and Latapy showed that the average proportion oscillates when p

grows. These oscillations was analysed in [Blondel et al. 2007]. The average number of edges (not the proportion) also oscillates when p grows (see Fig. 1).

In this paper we study *the number of vertices (or edges) which lie on all shortest paths between two vertices.* In this case also, we observe that the averages fluctuates when p varies. The average number of edges (Fig. 2(a)) oscillates in similar way to the average number of vertices (Fig. 2(b)). A sharp increase of average around $p = 2 * 10^{-3}$ corresponds to well-known phenomenon in Erdős–Rényi graphs: phase transition — birth of giant component (see [Bollobás 2001] for detailed explanation).

Surprisingly, we observe that the distribution is non-trivial, with several

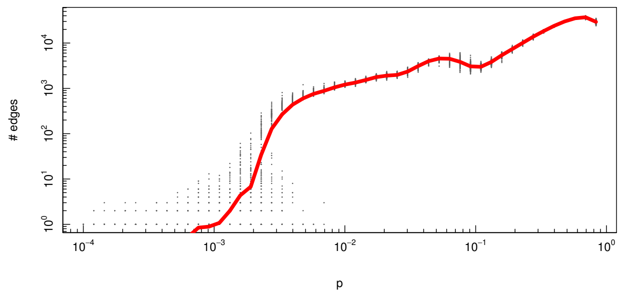
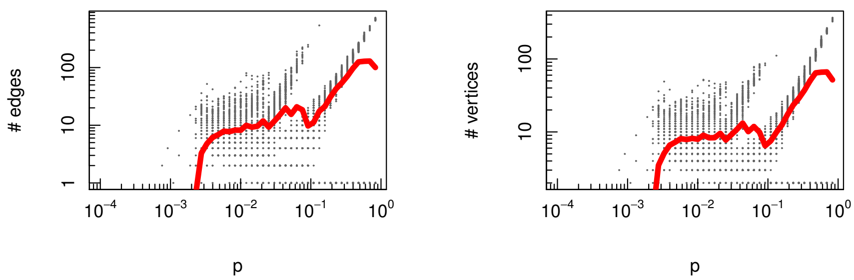


Figure 1: Evolution with p (log scale) of the number of edges (log scale) that lie on all shortest paths from a vertex to all other vertices in $G(n, p)$ with $n = 500$ vertices. Each grey point corresponds to an observed value. Red points represent the average over 200 graphs. A fast growth around $p = 2 * 10^{-3}$ corresponds to a phase transition.



(a) edges

(b) vertices

Figure 2: Evolution with p (log scale) of the number of edges and vertices (log scale) that lie on a shortest path between two vertices from $G(n, p)$ with $n = 500$ vertices. Each grey point corresponds to an observed value. Red points represent the average over 200 graphs. A fast growth around $p = 2 * 10^{-3}$ corresponds to a phase transition.

local maxima and that the average is a combination of these maxima and not a value that can be reached in all cases.

The rest of the paper is organised as follows. In Section 2 we give the necessary definitions. Next, in Section 3 we study complete and quasi-complete graphs (i.e. graphs obtained from complete graphs by removing an edge). The results from the Section 3, being a bit trivial, primarily serve to form an intuition about the number of vertices lie on all shortest paths between two chosen vertices. In Section 4 we consider dense random graphs and in Section 5 we study sparse random graphs with unbounded mean degree. In all cases (except the complete graphs) we find a non-trivial multimodal distribution of the number of vertices which lie on all shortest path between two vertices. We explain this phenomenon. Finally, we summarise our main results and discuss possible future works in Section 6.

2 Definitions

In this paper we are interested in the number of vertices or edges that belong to the shortest paths between two vertices. We give the necessary definitions below.

Definition 2.1: Given a graph and its two vertices u and v, let $SPS(u,v)$ be the subgraph of all shortest paths between u and v, i.e. $SPS(u,v)$ contains all edges and vertices that belong to all shortest paths between u and v.

Definition 2.2: Given a graph and its two vertices u and v, let $S(u,v)$ be the number of vertices in $SPS(u,v)$. When u and v are not connected $S(u,v)$ is equal to 0. Similarly, $S_E(u,v)$ will denote the number of edges in $SPS(u,v)$.

Definition 2.3: For a given graph, by $\text{đ}(u,v)$ we denote the distance between its two vertices u and v, i.e. the length of a shortest path from u to v. If between u and v there is no path at all, we say $\text{đ}(u,v) = \infty$. For brevity we write $\overset{x}{\widetilde{uv}}$ instead of $\text{đ}(u,v) = x$.

Definition 2.4: A random graph $G(n,p)$ is a graph with n vertices such that each edge between different vertices is included in the graph with probability p.

Fix two distinct vertices u and v from a given set V of n vertices, and consider all realisations of $G(n,p)$ over V as a probability space.

Definition 2.5: Let $f_d = \Pr\left[\text{đ}(u,v) = d\right]$ and $f_{>d} = \Pr\left[\text{đ}(u,v) > d\right]$.

Abusing the notation we will denote by S the random variable for the number of vertices in $SPS(u,v)$ when there is no ambiguity. Analogously, we use S_E for the random variable of the number of edges in $SPS(u,v)$.

3 Complete and quasi-complete graphs

Let us consider complete and quasi-complete graphs in order to form an intuition about the distribution of S. In the case of complete graph K_n the structure of $SPS(u,v)$ is trivial, because $SPS(u,v)$ contains only the vertices u, v and the edge uv. Consider now a quasi-complete graph $K_n - ab$, i.e. a graph obtained from K_n by removing an edge ab.

Proposition 3.1: *For any distinct vertices u and v from the quasi-complete graph $K_n - ab$ we have:*

$$S(u,v) = \begin{cases} n & \text{if } \{u,v\} = \{a,b\}\,, \\ 2 & \text{otherwise}\,. \end{cases}$$

Proof. Suppose $u = a$ and $v = b$ (or conversely). Vertices a and b are not directly connected, but they are connected by $n-2$ paths of length 2 (see for example Fig. 3). The union of these paths contains all vertices of our graph, so $S(a,b) = n$. Otherwise, u and v are directly connected, and $S(u,v) = 2$. □

The size distribution of shortest path subgraphs of the quasi-complete graph contains a valley $[3, n-1]$, and two peaks: 2, $n-1$.

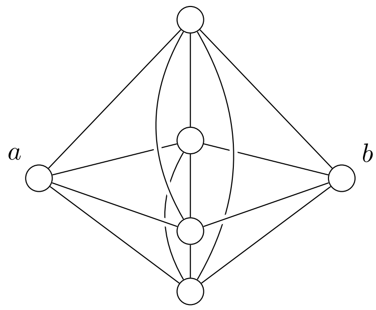


Figure 3: $K_6 - ab$. There are 4 different shortest paths between vertices a and b, while there is only 1 shortest path between any other two vertices.

4 Dense random graphs (p is fixed, $n \to \infty$)

In this section we study dense random graphs, i.e. graphs with constant density p. First, we show that such graphs have diameter 2. This allows us to consider only two cases: $\overset{1}{\widetilde{uv}}$ and $\overset{2}{\widetilde{uv}}$. We study these cases, and we show that the size distribution of $SPS(u,v)$ looks similar to the two-peak distribution from Proposition 3.1.

It is well known that almost all random graphs have diameter 2 (see for example [Moon and Moser 1966]). We present a similar result:

Theorem 4.1 (Random graphs with constant density have diameter 2)**:** *For any given $p > 0$ we have $f_1 = p$, $f_2 \to 1-p$, and $f_{>2} \to 0$ as $n \to \infty$.*

Proof. By the definition of $G(n,p)$ we have $f_1 = p$ and $f_{>1} = 1-p$. Next, the distance between two nonadjacent vertices u and v is larger than 2 if and only if they have no common neighbours. For a vertex different from u and v this happens with probability $1-p^2$, and it should be true for the $n-2$ remaining vertices distinct from u and v. So, we have:

$$f_{>2} = f_{>1}(1-p^2)^{n-2}\,.$$

Since $p > 0$ is fixed, $\lim_{n\to\infty} f_{>2} = 0$. □

Theorem 4.1 shows that in dense random graphs there is almost surely only two cases: $\overset{1}{\widetilde{uv}}$ and $\overset{2}{\widetilde{uv}}$. The structure of $SPS(u,v)$ in the former case is trivial, because $SPS(u,v)$ contains only the vertices u, v and the edge uv. In the rest of this section we study the later case. Denote by Y the number of vertices that are directly connected to both u and v.

Lemma 4.2: *Y is a binomial random variable with parameters $n-2$ and success probability p^2:*

$$Y \sim B(n-2, p^2)\,.$$

Proof. The probability that any vertex c is directly connected to both u and v is equal to p^2. We have $n-2$ vertices which are independently susceptible to lie between u and v. □

Theorem 4.3: *When the distance between u and v is equal to 2, the probability function* $\Pr\left[Y=k|\ \overset{2}{\widetilde{uv}}\right]$ *is equal to*

$$\begin{cases} 0 & \text{if } k=0\,, \\ \frac{\Pr[Y=k]}{1-(1-p^2)^{n-2}} & \text{if } k\geq 1\,. \end{cases}$$

Proof. From the definition of conditioned probability we have

$$\Pr\left[Y=k|\ \overset{2}{\widetilde{uv}}\right] = \frac{\Pr\left[Y=k \text{ and } \overset{2}{\widetilde{uv}}\right]}{\Pr\left[\overset{2}{\widetilde{uv}}\right]}\,.$$

Let A be an event "there is no edge between u and v". Observe that

$$\overset{2}{\widetilde{uv}} \Longleftrightarrow A \text{ and } Y \geq 1\,.$$

A is independent from Y and $\Pr[A] = 1-p$, so

$$\Pr\left[Y=k|\ \overset{2}{\widetilde{uv}}\right] = \frac{(1-p)\Pr\left[Y=k \text{ and } Y\geq 1\right]}{(1-p)(1-(1-p^2)^{n-2})}\,.$$

Note that

$$\Pr\left[Y=k \text{ and } Y\geq 1\right] = \begin{cases} 0 & \text{if } k=0\,, \\ \Pr[Y=k] & \text{if } k\geq 1\,. \end{cases}$$

The claimed formula easily follows. □

Corollary 4.4: *When the distance between u and v is equal to 2, we have the following expressions for the expectation and the variance of S:*

$$\begin{aligned} \mathbb{E}\left[S|\ \overset{2}{\widetilde{uv}}\right] &= 2 + \frac{(n-2)p^2}{1-(1-p^2)^{n-2}}\,, \\ \mathrm{Var}\left[S|\ \overset{2}{\widetilde{uv}}\right] &= \frac{(n-2)\,p^2\left(1-p^2+(n-2)p^2\right)}{1-(1-p^2)^{n-2}} - \left(\frac{(n-2)p^2}{1-(1-p^2)^{n-2}}\right)^2\,. \end{aligned}$$

Proof. When the distance between u and v is equals to 2, $SPS(u,v)$ contains $2+Y$ vertices. Thus, we have $\mathbb{E}\left[S|\overset{2}{\widetilde{uv}}\right] = 2 + \mathbb{E}\left[Y|\overset{2}{\widetilde{uv}}\right]$. From Lemma 4.2 we know that Y is a binomial random variable with parameters $n-2$ and success probability p^2. From Theorem 4.3 we know the probability function $\Pr\left[Y=k|\overset{2}{\widetilde{uv}}\right]$. Next, we write

$$\begin{aligned}
\mathbb{E}\left[Y|\overset{2}{\widetilde{uv}}\right] &= \sum_{k=0}^{\infty} k \Pr\left[Y=k|\overset{2}{\widetilde{uv}}\right] \\
&= \sum_{k=1}^{\infty} \frac{k \Pr[Y=k]}{1-(1-p^2)^{n-2}} \\
&= \frac{\mathbb{E}[Y]}{1-(1-p^2)^{n-2}} .
\end{aligned}$$

Now, the claimed formula for the expectation can be easily obtained.

Let's see what happens with the variance

$$\begin{aligned}
\operatorname{Var}\left[S|\overset{2}{\widetilde{uv}}\right] &= \operatorname{Var}\left[Y|\overset{2}{\widetilde{uv}}\right] \\
&= \mathbb{E}\left[Y^2|\overset{2}{\widetilde{uv}}\right] - \left(\mathbb{E}\left[Y^2|\overset{2}{\widetilde{uv}}\right]\right)^2
\end{aligned}$$

Note, that

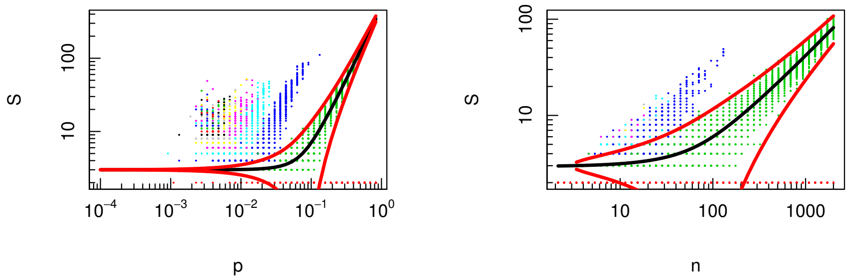
$$\begin{aligned}
\mathbb{E}\left[Y^2|\overset{2}{\widetilde{uv}}\right] &= \sum_{k=1}^{\infty} \frac{k^2 \Pr[Y=k]}{1-(1-p^2)^{n-2}} \\
&= \frac{\mathbb{E}[Y^2]}{1-(1-p^2)^{n-2}} \\
&= \frac{(n-2)\,p^2\left(1-p^2+(n-2)p^2\right)}{1-(1-p^2)^{n-2}} .
\end{aligned}$$

And finally

$$\operatorname{Var}\left[Y|\overset{2}{\widetilde{uv}}\right] = \frac{(n-2)\,p^2\left(1-p^2+(n-2)p^2\right)}{1-(1-p^2)^{n-2}} - \left(\frac{(n-2)p^2}{1-(1-p^2)^{n-2}}\right)^2$$

□

In order to illustrate Corollary 4.4 we performed some numerical simulations. Figure 4(a) shows the values of S in the case when n is fixed and $p \in [0,1]$. Figure 4(b) shows what happens when p is fixed but n grows. Different colours correspond to different distances between u and v. The black line represents $\mathbb{E}[S|\overset{2}{\widetilde{uv}}]$ and the red lines delimit the 89%-confidence interval. We see indeed that these simulations are in agreement with formulæ from Corollary 4.4.



(a) $n = 500, p \in [0, 1]$

(b) $n \in [2, 2000], p = 0.2$

Figure 4: Empirically measured values of S for different random graphs. Each point corresponds to an observed value of S. Different colours correspond to different distances between u and v (red: $\overset{1}{\widetilde{uv}}$, green: $\overset{2}{\widetilde{uv}}$, blue: $\overset{3}{\widetilde{uv}}$, cyan: $\overset{4}{\widetilde{uv}}$, magenta: $\overset{5}{\widetilde{uv}}$). The black line represents $\mathbb{E}[S|\ \overset{2}{\widetilde{uv}}]$ and the red lines delimit the 89%-confidence interval. A log scale is used for both axes.

Corollary 4.5: *When the distance between u and v is equal to 2, we have:*

$$\Pr\left[S_E = 2k|\ \overset{2}{\widetilde{uv}}\right] = \Pr\left[Y = k|\ \overset{2}{\widetilde{uv}}\right] .$$

Proof. It is sufficient to realise that for each vertex $c \in SPS(u,v), c \notin \{u, v\}$ there are exactly two distinct edges in $SPS(u,v)$, i.e. (c,u) and (c,v). $\square$

Finally, the probability mass function of S is a mixture of two functions: the first corresponds to the case $\overset{1}{\widetilde{uv}}$ and the second to $\overset{2}{\widetilde{uv}}$ (see Fig. 5):

$$S = \begin{cases} 2 & \text{if } \overset{1}{\widetilde{uv}}\,, \\ 2 + Y & \text{if } \overset{2}{\widetilde{uv}}\,, \text{ where } \Pr\left[Y = k|\ \overset{2}{\widetilde{uv}}\right] = \begin{cases} 0 & \text{if } k = 0\,, \\ \frac{\Pr[Y=k]}{1-(1-p^2)^{n-2}} & \text{if } k \geq 1\,. \end{cases} \end{cases}$$

The distribution of S has two local maxima, and the average size of $SPS(u,v)$ lies in the valley between these maxima. This finally explains why the real values of $S(u,v)$ and $S_E(u,v)$ are very different from the average.

5 Sparse random graphs with unbounded mean degree($p \to 0$ and $np \to \infty$ as $n \to \infty$)

We say that a random graph is sparse, when its density p tends to zero, as n goes to infinity. There are two classes of sparse graphs: (i) mean degree is constant ($np = c$), (ii) mean degree is unbounded ($np \to \infty$). Here we study the

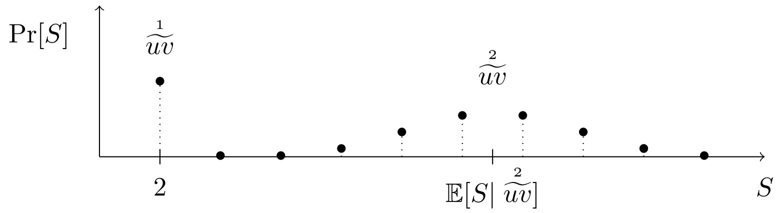


Figure 5: Schematic representation of the probability mass function of S.

second case. As Figure 4(a) suggests, when p decreases, the number of peaks in the distribution of S grows. For example, a third peak (dark-blue points on the Fig. 4(a)) appears when the probability that $d(u,v) = 3$ becomes non-negligible.

Let us give an intuitive explanation for the fact that there are several peaks and valleys in the size distribution of $SPS(u,v)$. We observe that $S(u,v) \geqslant d(u,v) + 1$, so, when $d(u,v)$ grows, $S(u,v)$ also grows. Intuitively, when our graphs are not like trees, $S(u,v)$ grows much faster than $d(u,v)$:

$$d(u,v) > d(u',v') \Rightarrow S(u,v) \gg S(u',v')\,.$$

Therefore, each observed value for $d(u,v)$ will correspond to a peak in the distribution of S.

In subsection 5.1 we give an approximation of the expected number of vertices in $SPS(u,v)$ in the case when we know the distance between u and v. In subsection 5.2 we give a classification of sparse graphs with unbounded mean degree, and we study the size distribution of $SPS(u,v)$ according to this classification.

5.1 Approximated expectation of the size of *SPS*

Recall that we denote by f_d (resp. $f_{>d}$) the probability that the distance between two vertices is equal to d (resp. greater than d). Authors in [Blondel et al. 2007] showed that f_d can be approximated by a recurrent formula:

$$\begin{aligned} f_{>0} &= 1 - \frac{1}{n}\,,\\ f_{>d} &= \big(1 - \frac{1}{n}\big)(1-p)^{(1-f_{>d-1})n}\,,\\ f_d &= f_{>d-1} - f_{>d}\,. \end{aligned}$$

We refer interested reader to [Blondel et al. 2007] for details about $f_{>d}$. Note however that the authors considered the case where the two chosen vertices are not necessarily distinct. Therefore, their definition of f_d is a bit different from our, but asymptotically they coincide.

Approximation 5.1: *When the distance between u and v is equal to x, we have the following approximation for the expectation of S.*

$$\mathbb{E}\left[S\middle|\ \overset{x}{\widetilde{uv}}\right] \approx x + 1 + (n-x-1)\sum_{y=1}^{x-1} f_y f_{x-y}\,.$$

Idea. First of all, it should be noted that the expectation of S, conditioned on the distance being x, is equal to

$$2 + (n-2) * \Pr\left[c \in SPS(u,v) |\ \overset{x}{\widetilde{uv}}\right].$$

But it seems difficult to calculate exact expectation, so we present here only an (over)approximation. When $\overset{x}{\widetilde{uv}}$, we know that $S \geqslant x+1$. There are $n-x-1$ possible vertices which also can lie on a shortest path between u and v, so we have

$$\mathbb{E}\left[S |\ \overset{x}{\widetilde{uv}}\right] \approx x + 1 + \Pr\left[c \in SPS(u,v) |\ \overset{x}{\widetilde{uv}}\right](n-x-1). \tag{1}$$

Note that a vertex c is on a shortest path between u and v if and only if $d(u,v) = d(u,c) + d(c,v)$. Therefore we have:

$$\Pr\left[c \in SPS(u,v) |\ \overset{x}{\widetilde{uv}}\right] = \frac{\sum_{y=1}^{x-1} \Pr\left[\overset{y}{\widetilde{uc}} \text{ and } \overset{x-y}{\widetilde{cv}} \text{ and } \overset{x}{\widetilde{uv}}\right]}{\Pr\left[\overset{x}{\widetilde{uv}}\right]}$$

Assuming that events $\overset{y}{\widetilde{uc}}$, $\overset{x-y}{\widetilde{cv}}$ and $\overset{x}{\widetilde{uv}}$ are mutually independent and identically distributed (actually it is not true, because there is triangular inequality that creates some dependencies), we approximate:

$$\Pr\left[\overset{y}{\widetilde{uc}} \text{ and } \overset{x-y}{\widetilde{cv}}\right] \approx \Pr\left[\overset{y}{\widetilde{uc}}\right]\Pr\left[\overset{x-y}{\widetilde{cv}}\right] \approx f_y f_{x-y}\,,$$

and finally

$$\Pr\left[c \in SPS(u,v) |\ \overset{x}{\widetilde{uv}}\right] \approx \sum_{y=1}^{x-1} f_y f_{x-y}\,. \tag{2}$$

Using Blondel's relation and our formulæ (1) and (2), we are able to obtain the claimed approximation.

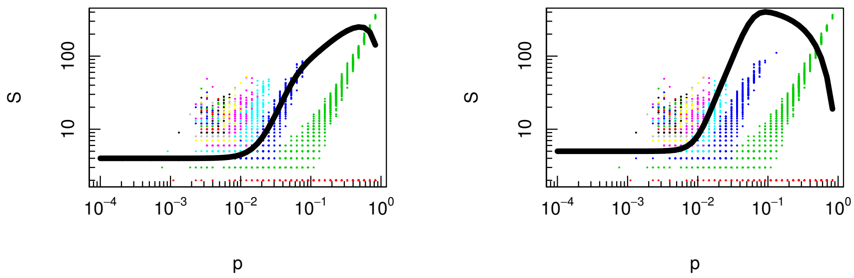
Figure 6 illustrates our approximation when $d(u,v) \in \{3,4\}$. Clearly we see several peaks, i.e. typical values of S. Our approximation corresponds to the centres of these peaks. There are valleys between peaks. But these valleys vanish when p is very small, due to the variance of S.

Typically our approximation gives very good estimations compared to experimental data. However, when $d(u,v)$ is very large (compared to the average distance), the result of approximation is slightly inadequate. This happens due to the following reasons: (i) Blondel *et al.* expression for f_d is not exact; (ii) we neglect the dependence between some events.

Also when our graph is very dense, f_3 and f_4 are almost zero (see Theorem 4.1), and there are no points around the "right tails" of black curves.

5.2 Classification of sparse graphs according to the distance distribution

In appendix of [Blondel et al. 2007] the following was shown:



(a) $\overset{3}{\widetilde{uv}}$ (b) $\overset{4}{\widetilde{uv}}$

Figure 6: Empirically measured values of S in random graph with 500 vertices. Different colours correspond to different distances between u and v (red: $\overset{1}{\widetilde{uv}}$, green: $\overset{2}{\widetilde{uv}}$, blue: $\overset{3}{\widetilde{uv}}$, cyan: $\overset{4}{\widetilde{uv}}$, magenta: $\overset{5}{\widetilde{uv}}$). Approximated $\mathbb{E}\left[S|\ \overset{x}{\widetilde{uv}}\right]$ is represented by black lines. A log scale is used for both axes.

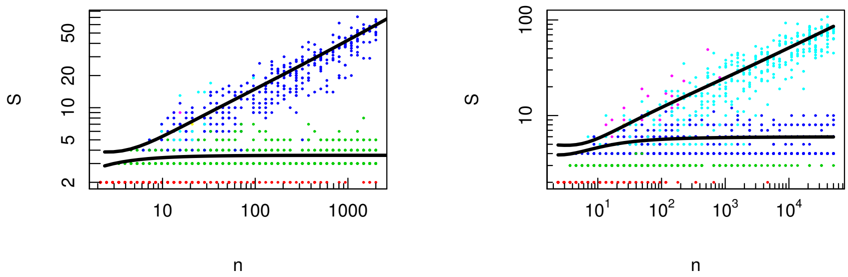
Theorem 5.2: *For any given $d \geqslant 2$ and any given $\lambda \in (0, \infty)$, if $n^{d-1}p^d = \lambda$ we have:*

$$\begin{aligned}\lim_{n\to\infty} f_{<d} &= 0\,,\\ \lim_{n\to\infty} f_{d} &= 1 - e^{-\lambda}\,,\\ \lim_{n\to\infty} f_{d+1} &= e^{-\lambda}\,,\\ \lim_{n\to\infty} f_{>d+1} &= 0\,.\end{aligned}$$

Informally, this means that when $p = \sqrt[d]{\frac{\lambda}{n^{d-1}}}$ and n is sufficiently large, there are only two possibilities: $\overset{d}{\widetilde{uv}}$ and $\overset{d+1}{\widetilde{uv}}$. It can be shown that $\lim_{n\to\infty} E\left[S|\ \overset{d}{\widetilde{uv}}\ \right]$ exists and depends only on λ, while $E\left[S|\ \overset{d+1}{\widetilde{uv}}\ \right]$ grows with n.

For the random graphs with constant density, we have: $d = 1$, $\mathbb{E}[S|\ \overset{1}{\widetilde{uv}}] = 2$ and $\mathbb{E}\left[S_E|\ \overset{2}{\widetilde{uv}}\right] \approx np^2$ (see Proposition 4.3 and Corollary 4.4). Therefore, we have similar size distributions of $SPS(u, v)$ in the case of dense and sparse random graphs.

We performed numerical simulations using two families of sparse random graphs: (A) $np^2 = 1$ and (B) $n^2p^3 = 1$. For each value of p we generated 200 random graphs $G(n, p)$ and we measured S (each point in Fig. 7 corresponds to a measured value of S, different colours correspond to different distances between vertices). Finally, we let n go to infinity. We see that $E\left[S|\ \overset{d}{\widetilde{uv}}\ \right]$ stabilises around some value, but $E\left[S|\ \overset{d+1}{\widetilde{uv}}\ \right]$ grows unboundedly. The valley between these two typical values of S grows also with n (see Fig. 7).



(a) $np^2 = 1$ (b) $n^2p^3 = 1$

Figure 7: Empirically measured values of S. Different colours correspond to different distances between u and v. Our approximation for $\mathbb{E}\big[S|\overset{x}{\widetilde{uv}}\,\big]$ is represented by black lines (on the left: $x \in \{2, 3\}$, on the right: $x \in \{3, 4\}$). A log scale is used for both axes.

6 Summary and discussions

In this paper we study the size of shortest path subgraph between two vertices. We denote by S the number of vertices in that subgraph. When we consider a family of random graphs $G(n, p)$, S becomes a random variable. This paper results in a characterisation of S, see Theorem 4.3, Corollary 4.4 and Approximation 5.1.

The probability mass function of S has several local maxima (peaks). Each peak corresponds to a possible distance between u and v. Between such peaks we have valleys of "improbable" sizes of SPS, in other words the distribution of S is multimodal.

The structure of $SPS(u, v)$ in the case of $\overset{1}{\widetilde{uv}}$ is trivial. We give the exact distribution for the size of $SPS(u, v)$ (in terms of edges and vertices) in the case of $\overset{2}{\widetilde{uv}}$ (see Theorem 4.3, Corollary 4.4 and Corollary 4.5). For other cases (e.g. $\overset{3}{\widetilde{uv}}$, $\overset{4}{\widetilde{uv}}$) we have an approximated representation of expected number of vertices in $SPS(u, v)$ (see Approximation 5.1). Better approximations (or even exact distributions) are parts of a future work. Another part consists in studying real-world networks or other models of random graphs (e.g. power-law graphs). Future works may also investigate another important class of sparse random graph, when mean degree is constant ($np \to c > 1$), using the methodology described in [H. van den Esker et al. 2008].

Our study of the size of $SPS(u, v)$ in the case of random graphs gives some insights about $SPS(u, v)$ in real-world networks (e.g. about the network structure between two computers in the Internet). Moreover, the notion of $SPS(u, v)$ can be considered as a similarity measure between vertices u and v, that may be helpful in community detection.

Acknowledgements

Authors are grateful to Michèle Soria, Jean-Loup Guillaume and anonymous reviewer for helpful comments and proof-readings. Authors also thank Maximilien Danisch and Daniel Bernardès for helpful discussions. This work was partly funded by a grant from the *Agence Nationale de la Recherche*, with reference ANR-10-JCJC-0202. It was also supported in part by the European Commission through the FP7 FIRE project EULER (Grant No.258307) and by the EIT-ICT Labs Action Line Future Networking Solutions (FNS).

References

Blondel, V. D., Guillaume, J.-L., Hendrickx, J. M., and Jungers, R. M. (2007). Distance distribution in random graphs and application to network exploration. *Phys. Rev. E*, 76:066–101.

Bollobás, B. (2001). *Random Graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen*, 6:290.

Gilbert, E. N. (1959). Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144.

Guillaume, J.-L. and Latapy, M. (2005). Complex network metrology. *Complex Systems*, (16):83–94.

H. van den Esker, R. van der Hofstad, and G. Hooghiemstra (2008). Universality for the distance in finite variance random graphs. *J. Stat. Phys.*, 133(1):169–202.

Latapy, M., Magnien, C., and Ouédraogo, F. (2011). A radar for the internet. *Complex Systems*, 20(1):23–30.

Magnien, C., Medem, A., Kirgizov, S., and Tarissan, F. (2013). Towards realistic modeling of IP-level routing topology dynamics. *Networking Science*, pages 1–10.

Medem, A., Magnien, C., and Tarissan, F. (2012). Impact of power-law topology on IP-level routing dynamics: simulation results. In *Proceedings of the Fourth International Workshop on Network Science for Communication Networks (NetSciCom)*.

Moon, J. W. and Moser, L. (1966). Almost all (0,1) matrices are primitive. *Studia Scientiarum Mathematicarum Hungarica*, 1:153–156.

Rapoport, A. (1948). Cycle distributions in random nets. *The bulletin of mathematical biophysics*, 10:145–157.

Solomonoff, R. and Rapoport, A. (1951). Connectivity of random nets. *Bulletin of Mathematical Biology*, 13(2):107–117.