UNIVERSITY OF FRIBOURG

MASTER PROJECT

Giant connected component in networks

Author:
Benoît RICHARD

Supervisors:
Dr. Guiyuan SHI
Prof. Yi-Cheng ZHANG

Theoretical Interdisciplinary Physics Group
Department of Physics

May 16, 2018

University of Fribourg

Abstract

Faculty of Science
Department of Physics

Master Project

Giant connected component in networks

by Benoît RICHARD

Write the abstract

Contents

List of Symbols

Symbol	Description	Math. definition
$\deg(v)$	Degree of vertex v	
$g_0(z)$	Generating function for the degree of uniformly chosen nodes	$\sum_{k=0}^{\infty} p_k z^k$
$g_1(z)$	Generating function for the degree of nodes reached by following an edge	$\sum_{k=0}^{\infty} q_k z^k$
n	Number of nodes in the network	
$N(v)$	Neighborhood of a vertex v	
p_k	Probability that a random node has degree k	$P_0(\deg(v) = k)$
P_0	Probability starting from a uniformly chosen node	
P_1	Probability starting from a node reached by following an edge	
q_k	Probability that a node reached by following a edge has degree k	$P_1(\deg(v) = k)$
S	Fraction of the network which is part of the GCC in the large n limit	$P_0(v \in GCC)$
u	Probability that a node reached by following an edge from is not part of the GCC	$P_1(v \notin GCC)$
$g_0^{(i)}(z)$	Generating function for the degree of uniformly chosen nodes in layer i	
$g_1^{(i)}(z)$	Generating function for the degree of nodes reached by following an edge in layer i	
$p_k^{(i)}$	Probability that a uniformly chosen node has degree k in layer i	

Todo list

Chapter 1

Introduction

Many systems in real world can be conceptually represented as objects being connected to others. Such representation is called a network. For example, a road network can be represented as a set of crossings that are connected by direct roads. However, the concept of network does not require the object or the link between them to be physical. We can represent friendship relations as a network: two people are connected if they consider to be friends.

Mathematically, networks are represented as *graphs*. A graph is an object composed of a set V of nodes (also referred to as vertices) and a set of edges E. An edge is characterized by the fact that it connects two nodes together, which in mathematical terms translate to the fact that an edge can be written as a pair of nodes or equivalently $E \subset \{(v_1, v_2) | v_1, v_2 \in V\}$. To fit the numerous systems many extension of this simple model can be considered, for example edges may have a direction (*directed graph*), meaning that $(v_1, v_2) \neq (v_2, v_1)$, edges or vertices can also have carry a value (*weighted graph*) or multiple edges between two vertices may be allowed.

Introduce generating functions

Chapter 2

Single layer networks

Single layer networks correspond to the classic picture of a network.

2.1 Configuration model

Rather than studying peculiar and unique networks, we would like to find properties of whole classes of networks. This allows to average properties over such network classes and thus may allow to identify characteristics fundamental to the network structure.

Since we would like to consider classes of network, we must first classify them. A common and usefulway of doing so is to consider that two networks are element of the same class Ω if they have the same set of degrees. This definition is useful for two main reasons. First, the set of degrees of a network is easily measured if its structure is known. Second it is possible to generate a random network uniformly chosen in its class using the so called *configuration model*. The main goal of this section is to describe the configuration model and how it achieves this.

add ref

Consider a network with n vertices with degrees d_i, $i = 1 \dots n$. If we cut every edges in two then every vertex will be disconnected from the other and keep a number of "half edges" equal to its degree. In the context of the configuration model, we call such "half edge" a *stub*. The resulting set of vertices and stub is independent of the network structure, but common for all networks with the same degree distribution. The idea of the configuration model is thus to start from this state and bond stubs two by two in a meaningful way to recreate edges of the network.

It appears that a good way to choose which stubs to bond together is to uniformly pick two amongst all of them. This procedure is good in the sense that it generate a network uniformly chosen in the class Ω of the network.

Network distribution and multi- and self-edges distribution

2.2 Giant connected component

An interesting property of a network is the presence and size of *connected component*. A set of nodes is said to be connected if there is a path of edges from any of its node to any other. All networks can divided in connected components such that all nodes are element of exactly one component, as can be seen on fig. 2.1. The connectedness of network is crucial in many real world realisations of networks. In particular any logistic network, such as power grid networks, rail road network or the internet network, is functional only if it is able to transfer goods or services (electricity, passengers or informations) from any node to any other.

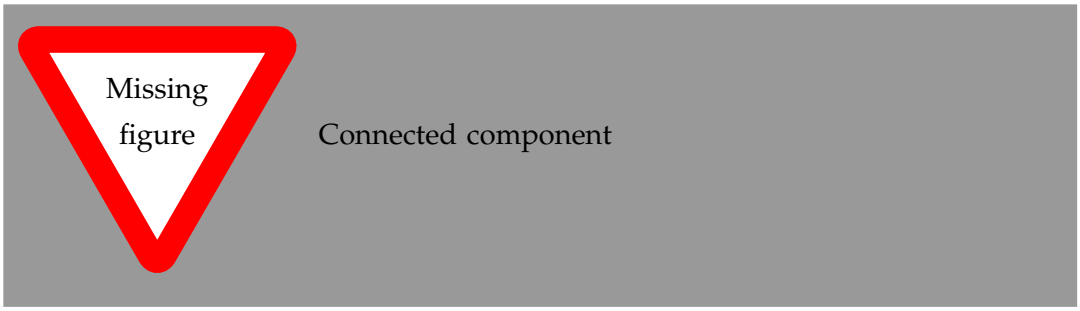


FIGURE 2.1

As we will see in section 2.3, it is really hard to generate a fully connected network chosen uniformly in its class. A simpler, yet powerful approach, is to instead consider the biggest connected part of a network generated using the configuration model. This component, if its relative size does not vanish in the limit of large n, is called the *giant connected component* (GCC). The first question is: what will be the size of the giant connected component ?

To determine this we first define u the probability that a node reached by following an edge is not part of the GCC. We can therefore write the probability S that a randomly chosen vertex is part of the GCC as

$$S = 1 - P_0(w \notin GCC\ \forall w \in N(v)) \tag{2.1}$$
$$= 1 - \sum_{k=0}^{\infty} P_0(w \notin GCC\ \forall w \in N(v) | \deg(v) = k) P_0(\deg(v) = k) \tag{2.2}$$
$$= 1 - \sum_{k=0}^{\infty} [P_1(w \notin GCC)]^k\, p_k \tag{2.3}$$
$$= 1 - \sum_{k=0}^{\infty} u^k p_k \tag{2.4}$$
$$= 1 - g_0(u). \tag{2.5}$$

We have now a compact expression for S in terms of u and the generating function of the degree distribution g_0. To determine u we notice that if a vertex is not part of the GCC, none of its neighbors is. This allows to write

$$u = P_1(w \notin GCC\ \forall w \in N(v)) \tag{2.6}$$
$$= \sum_{k=0}^{\infty} P_1(w \notin GCC\ \forall w \in N(v) | \deg(v) = k) P_1(\deg(v) = k) \tag{2.7}$$
$$= \sum_{k=0}^{\infty} u^k q_k \tag{2.8}$$
$$= g_1(u). \tag{2.9}$$

We end up with two equations to describe the GCC size

{Single layer GCC final}
$$S = 1 - g_0(u) \tag{2.10}$$
{Single layer u final}
$$u = g_1(u). \tag{2.11}$$

If we can solve the second one we immediately get the GCC size. However eq.

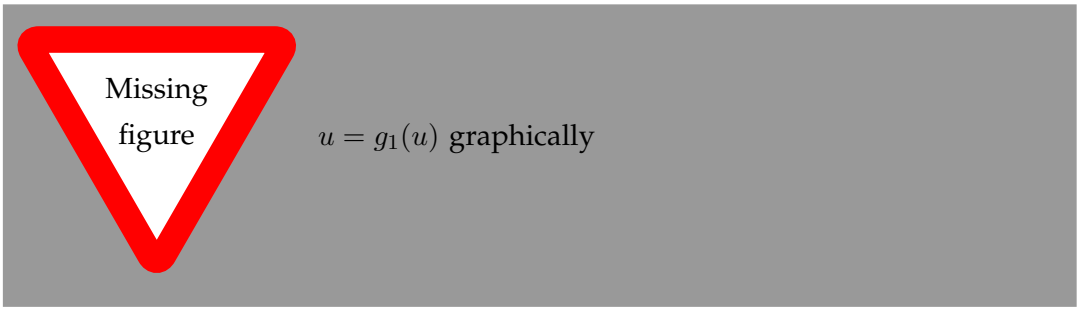


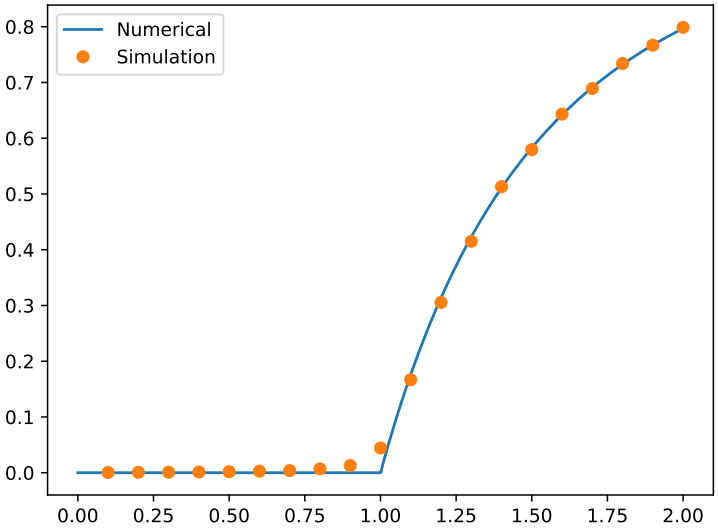
FIGURE 2.2

(2.10) only gives u implicitly and its form strongly depends on the degree distribution, therefore no general analytical solutions can be given. A particular solution is however always present for $u = 1$ as by definition $g_1(1) = 1$. This implies $S = 0$ and thus the absence of GCC.

Add reference to definition of g1

Add discussion/proof of existence of GCC when multiple solutions are present ?

(A)



(B)

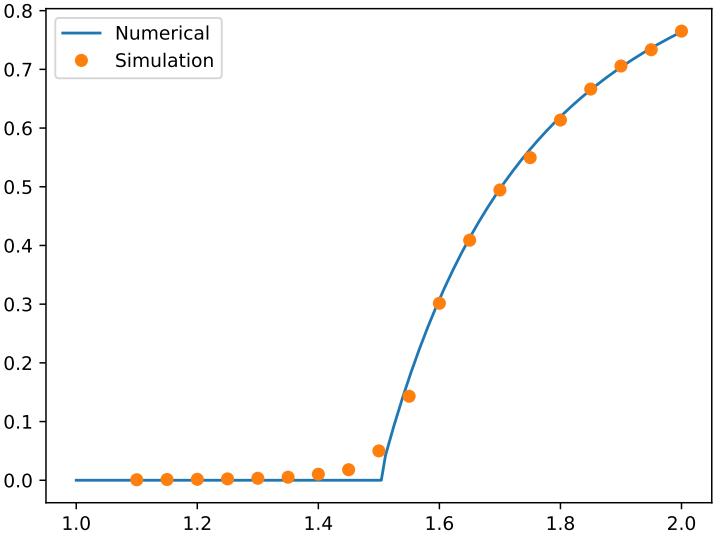


FIGURE 2.3: Numerical solution of eqs. (2.10) and (2.11), together with results on simulated networks. Results of simulations are average over 10 runs and the network size was set to 10^4 nodes. Simulations ran for several seconds in total, indicating that much larger network should be easily usable. (A) Poisson degree distribution. (B) Geometric degree distribution.

low degree saturation in GCC

2.3 Degree distribution in the GCC

Per Bayes theorem we have for two event A and B

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(B|A)\frac{P(A)}{P(B)}. \tag{2.12}$$

{Bayes theorem}

This allows us to compute the degree distribution of the vertices in the giant connected component

$$P(deg(v) = k|v \in GCC) = P(v \in GCC|deg(v) = k)\frac{P(deg(v) = k)}{P(v \in GCC)} \tag{2.13}$$

$$= (1 - P(v \notin GCC|deg(v) = k))\frac{p_k}{S} \tag{2.14}$$

$$= \frac{p_k}{S}(1 - u^k). \tag{2.15}$$

{Degree distribution in GCC}

By multiplying this expression by z^k for each k and summing, we find the generating function $G_0(z)$ of the degree distribution in the giant connected component

$$G_0(z) = \frac{1}{S}(g_0(z) - g_0(uz)). \tag{2.16}$$

Chapter 3

Multiplex network

3.1 Giant viable cluster

Consider a multiplex network with L layers. Let $g_0^{(i)}$ and $g_1^{(i)}$ be the generating functions of respectively the degree and the excess degree in layer i. Moreover define u_i as the probability that a vertex reached after following an edge in layer i is not part of the giant viable cluster. Then if we pick a vertex v at random the probability S that it is part of the giant viable cluster can be written as

$$S = P_0\left(\bigcap_{i=1}^{L} \exists w \in N_i(v)\ w \in GVC\right). \tag{3.1}$$

By requiring that the layers are independent from one others, we can rewrite S as a product

$$\begin{aligned}
S &= \prod_{i=1}^{L} P_0\left(\exists w \in N_i(v)\ w \in GVC\right) && (3.2)\\
&= \prod_{i=1}^{L} \left[1 - P\left(w \notin GVC\ \forall w \in N_i(v)\right)\right] && (3.3)\\
&= \prod_{i=1}^{L} \left[1 - \sum_{k=0}^{\infty} P\left((w \notin GVC\ \forall w \in N_i(v) | deg(v) = k\right) p_k^{(i)}\right] && (3.4)\\
&= \prod_{i=1}^{L} \left[1 - \sum_{k=0}^{\infty} u_i^k p_k^{(i)}\right] && (3.5)\\
&= \prod_{i=1}^{L} \left[1 - g_0^{(i)}(u_i)\right]. && (3.6)
\end{aligned}$$

{Multiplex GCC size final}

We can find u_j by a similar reasoning. First note that $1 - u_j$ is the probability that a vertex reached by following an edge in layer j is in the giant viable cluster. Which as before can be written in the form

$$\begin{aligned}
1 - u_j &= P_1^{(j)}\left(\bigcap_{i=1}^{L} \exists w \in N_i(v)\ w \in GVC\right) && (3.7)\\
&= \prod_{i=1}^{L} P_1^{(j)}\left(\exists w \in N_i(v)\ w \in GVC\right). && (3.8)
\end{aligned}$$

Since the layers are independent, the fact that we reached v by following an edge

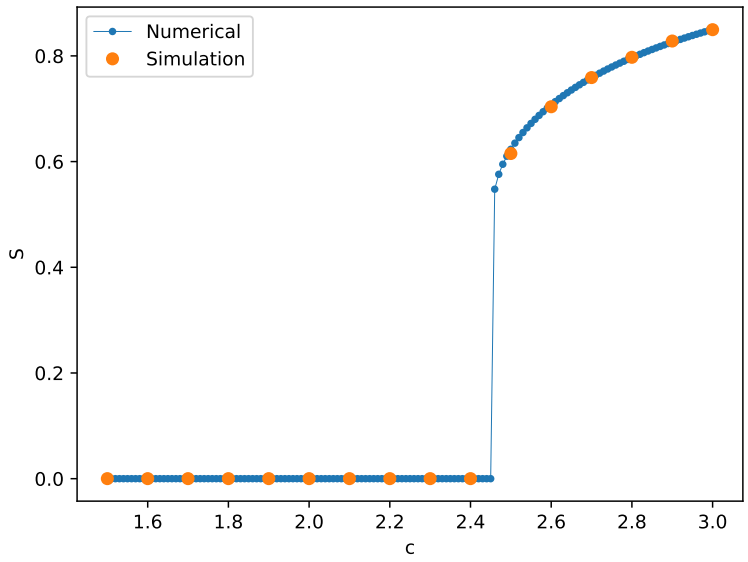
in layer j to reach vertex v is irrelevant in all other layers. However in layer j this means that the degree distribution follows the distribution $q_k^{(j)}$ rather than $p_k^{(j)}$. Putting this together we get

$$1 - u_j = \left[1 - \sum_{k=0}^{\infty} u_j^k q_k^{(j)}\right] \prod_{\substack{i=1 \\ i \neq j}}^{L} \left[1 - \sum_{k=0}^{\infty} u_i^k p_k^{(i)}\right] \tag{3.9}$$

{Multiplex u final}

$$= \left[1 - g_1^{(j)}(u_j)\right] \prod_{\substack{i=1 \\ i \neq j}}^{L} \left[1 - g_0^{(i)}(u_i)\right]. \tag{3.10}$$

(A)



(B)

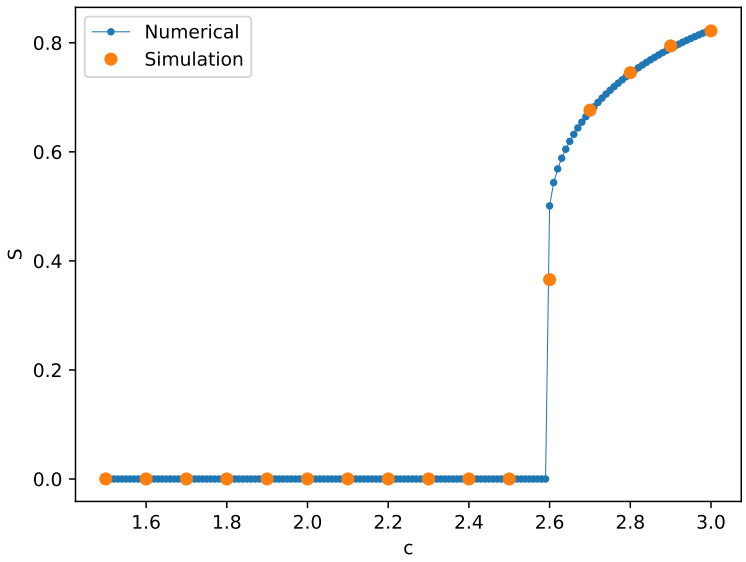


FIGURE 3.1: Numerical solution of eqs. (3.6) and (3.10), together with results on simulated networks for multiplex networks composed of two layers with the same distribution and mean number of edge c. Results of simulations are average over 10 runs and the network size was set to 10^4 nodes. Simulations ran for several seconds in total, indicating that much larger network should be easily usable. (A) Poisson degree distribution. (B) Geometric degree distribution.

3.2 Boundary condition

Equations (3.6) and (3.10) are in principle sufficient to determine the size S of the giant viable cluster in a multiplex network. We see that there always exists a trivial solution

$$u_j = 1, \qquad \forall j, \tag{3.11}$$
$$S = 0. \tag{3.12}$$

First notice that S can be zero if and only if all u_j are one. Indeed since $g_0^{(j)}$ is a strictly increasing function, we have

$$g_0^{(j)}(z) = 0 \quad \Leftrightarrow \quad z = 1. \tag{3.13}$$

So if $S = 0$, there exists k such that $u_k = 1$. If we put it back in eq. (3.10), it forces $1 - u_j = 0$ for all j, and thus all u_j are one.

Despite the fact that there is always an unique solution such that there is no giant viable cluster, the condition to have more than this solution is not immediately clear. To find it, observe that a new solution for eq. (3.10) appears when the curves

$$1 - u_j \tag{3.14}$$
$$\text{and} \qquad \left[1 - g_1^{(j)}(u_j)\right] \prod_{\substack{i=1 \\ i\neq j}}^{L} \left[1 - g_0^{(i)}(u_i)\right] \tag{3.15}$$

are tangent. So the derivative of both with respect to u_j must be equals, which means

$$-1 = \frac{\mathrm{d}}{\mathrm{d}u_j}(1 - u_j) = \frac{\mathrm{d}}{\mathrm{d}u_j}\left[1 - g_1^{(j)}(u_j)\right] \prod_{\substack{i=1 \\ i\neq j}}^{L} \left[1 - g_0^{(i)}(u_i)\right] \tag{3.16}$$
$$= -\left(\frac{\mathrm{d}}{\mathrm{d}u_j} g_1^{(j)}(u_j)\right) \prod_{\substack{i=1 \\ i\neq j}}^{L} \left[1 - g_0^{(i)}(u_i)\right] \tag{3.17}$$
$$- \sum_{k=1}^{L} \left[1 - g_1^{(j)}(u_j)\right] \frac{\mathrm{d}u_k}{\mathrm{d}u_j} \left(\frac{\mathrm{d}}{\mathrm{d}u_k} g_0^{(k)}(u_k)\right) \prod_{\substack{i=1 \\ i\neq j \\ i\neq k}}^{L} \left[1 - g_0^{(i)}(u_i)\right]. \tag{3.18}$$

This can be rewritten as

$$0 = \sum_{k=1}^{L} Q_{jk} \frac{\mathrm{d}u_k}{\mathrm{d}u_j} \tag{3.19}$$
$$= Q_{jj} + \sum_{\substack{k=1 \\ k\neq j}}^{L} Q_{jk} \frac{\mathrm{d}u_k}{\mathrm{d}\lambda} \frac{\mathrm{d}\lambda}{\mathrm{d}u_j}, \tag{3.20}$$

where

$$Q_{jj} = \left(\frac{\mathrm{d}}{\mathrm{d}u_j} g_1^{(j)}(u_j)\right) \prod_{\substack{i=1 \\ i\neq j}}^{L} \left[1 - g_0^{(i)}(u_i)\right] - 1, \tag{3.21}$$

$$Q_{jk} = \left[1 - g_1^{(j)}(u_j)\right] \left(\frac{\mathrm{d}}{\mathrm{d}u_k} g_0^{(k)}(u_k)\right) \prod_{\substack{i=1 \\ i\neq j \\ i\neq k}}^{L} \left[1 - g_0^{(i)}(u_i)\right], \qquad k \neq j, \tag{3.22}$$

and where λ is some parameter, from which the u_j depend. For example it could be a composition of the parameters of the degree distribution for each layer.

If u_j is a smooth enough function of λ and both are not independent we have

$$\frac{\mathrm{d}u_j}{\mathrm{d}\lambda} = \left(\frac{\mathrm{d}\lambda}{\mathrm{d}u_j}\right)^{-1} \neq 0 \tag{3.23}$$

and we can rewrite the tangency condition for each j as

$$\sum_{k=1}^{L} Q_{jk} \frac{\mathrm{d}u_k}{\mathrm{d}\lambda} = 0, \tag{3.24}$$

which can be translated in matrix form as

$$Q \cdot v = 0, \tag{3.25}$$

where Q is the matrix with elements Q_{kj} and v the vector with elements

$$v_k = \frac{\mathrm{d}u_k}{\mathrm{d}\lambda}. \tag{3.26}$$

The system has an obvious solution $v_k = 0$ for all k, but this implies that λ has been chosen independent of the u_k, a possibility which is not very interesting. Otherwise, the system has a solution if Q is singular, which implies

$$\det Q = 0. \tag{3.27}$$

{Tangency condition as determinant

In this case, the value of v is of not relevant to the problem, nor is the value of parameter λ.

This gives us an additional equation which allows in principle to find the boundary of the region (in the space of parameters of the degree distributions) where S is not zero. Indeed the set of equations (3.10) give us implicitly the u_j as a function of the degree distribution parameters, and eq. (3.27) restrict the parameters to where they first encounter a non trivial solution.