Random Hypergraphs and their applications

Study of topology in tagged social networks

G. Ghoshal[1] M.E.J. Newman[1] V. Zlatić[2]
G. Caldarelli[2,3]

[1]Department of Physics and Michigan Center for Theoretical Physics, University of Michigan, Ann Arbor Michigan USA

[2]**CNR Istituto dei Sistemi Complessi, University "Sapienza" Rome Italy**

[3]**London Institute for Mathematical Sciences, London Mayfair (visiting)**

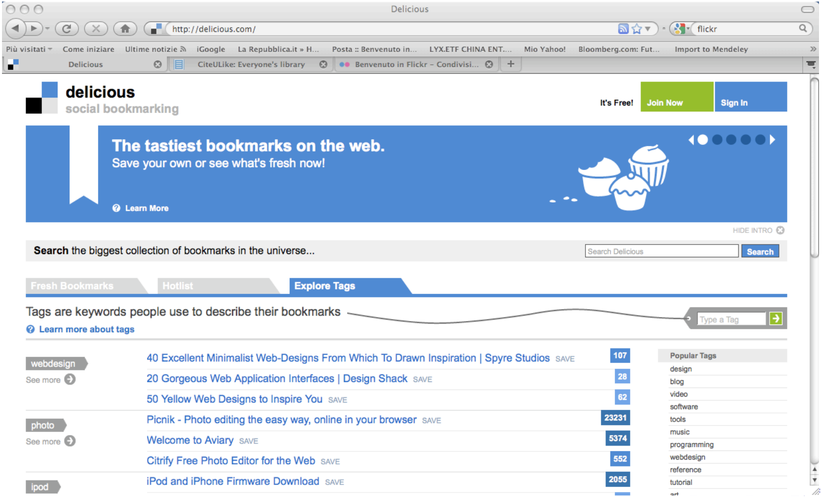
CABDyN Oxford January 18th 2011
www.guidocaldarelli.com/CABDyN.pdf

Overview

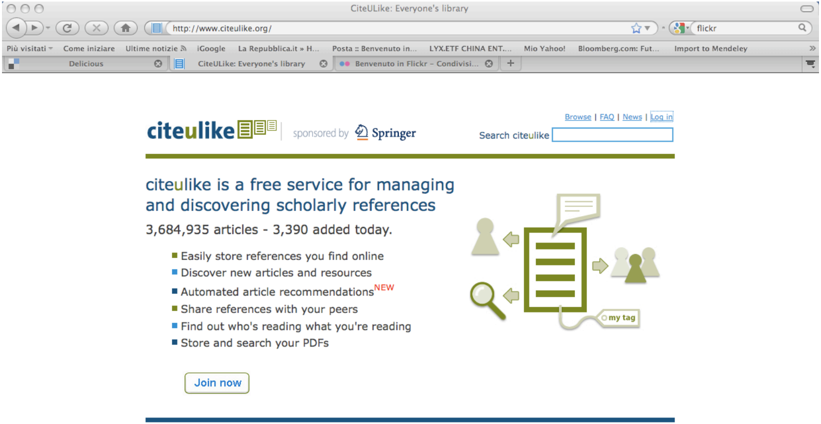
- I want to present a way to describe objects more complicated than networks
- This applies to
  - "tagged" (essentially social) networks
  - interacting networks
  - interconnected networks
- This is done with a generalization of graphs known as "hypergraphs" [1], [2]

[1] G. Ghoshal, V. Zlatić, GC, M.E.J. Newman *PRE* **80** 036118 (2009)
[2] G. Ghoshal, V. Zlatić, GC *PRE* **79** 066118 (2009)

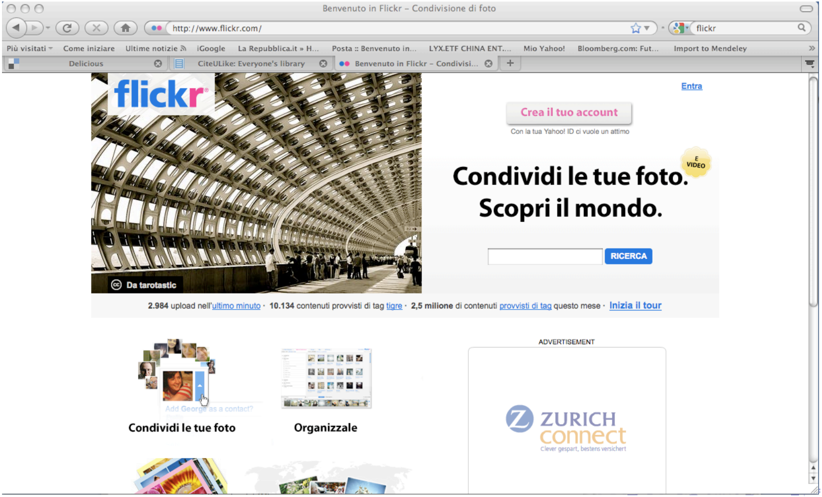
Delicious site



Citeulike site

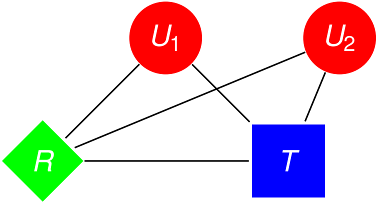


Flickr site


Benvenuto in Flickr – Condivisione di foto
http://www.flickr.com/
flickr
Entra
Crea il tuo account
Con la tua Yahoo! ID ci vuole un attimo
E VIDEO
Condividi le tue foto.
Scopri il mondo.
RICERCA
Da tarotastic
2.984 upload nell'ultimo minuto · 10.134 contenuti provvisti di tag tigre · 2,5 milione di contenuti provvisti di tag questo mese · Inizia il tour
ADVERTISEMENT
Condividi le tue foto
Organizzale
ZURICH connect
Clever gespart, bestens versichert

Think Hypergraph!

Hypergraph describe these systems in a compact way.

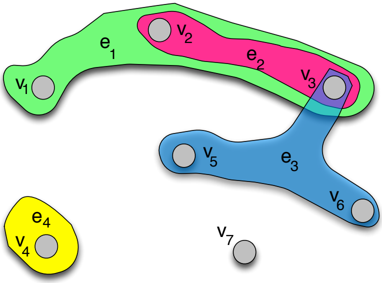


Hypergraph Basic Unit

The typical structure that you have in these systems is a triple

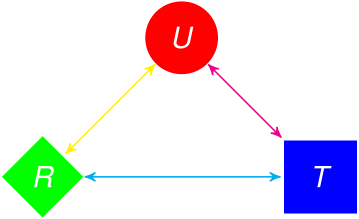
- A red vertex for the **user** (people)
- A green vertex for the **resource** (paper, picture)
- A blue vertex for the **tag** ("Graphs", "vacation" etc)

Hypergraph theory



- Hypergraphs are generalization of graphs,
- Hyperedges are arbitrary set of vertices
- Tagged systems are *3-uniform hypergraphs*

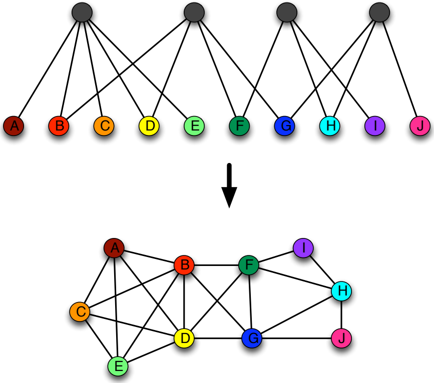
Hypergraphs and hyperedges



Following color code we have

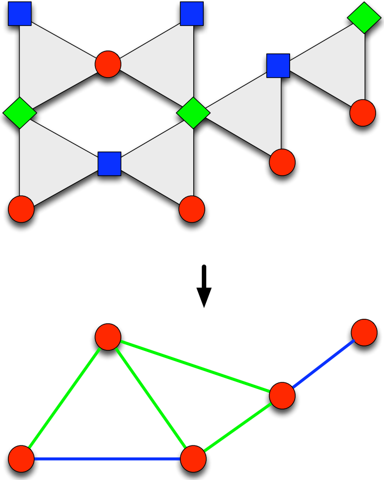
- U-R yellow edge between a red vertex and a green vertex
- U-T magenta edge between a red vertex and a blue vertex
- R-T cyan edge between a green vertex and a blue vertex

Projections



One approach is“to project" this triple structure along one of the three components.
Similarly to bipartite graphs of collaborations for ordinary graphs.

Hypergraph Projections



Hypergraphs can also be projected, but it is more interesting to consider them as a whole

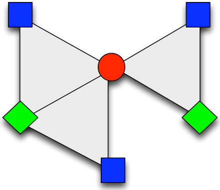
We define here:

- *degree* as the number of hyperedges neighbour
- *distance* as the number of hyperedges to travel
- *clustering* as the triples of hyperedges
- *communities* by considering the set of common hyperedges between vertices

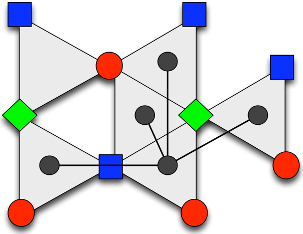
Degrees

Two generalizations are possible from Graph Theory

for a vertex/edge we count the hyperedges it participates in.



for an hyperedge we count the hyperedges neighbours



Expected values

The mean degree $\langle k_r \rangle$ of a red vertex in our network is given by the number of hyperedges in the network divided by the number of red vertices, and similarly for green and blue:

$$\langle k_r \rangle = \frac{M}{n_r}, \qquad \langle k_g \rangle = \frac{M}{n_g}, \qquad \langle k_b \rangle = \frac{M}{n_b}.$$

Rearranging these equations we can write:

$$n_r \langle k_r \rangle = n_g \langle k_g \rangle = n_b \langle k_b \rangle = M.$$

Thus the mean degrees of the different vertex types cannot be chosen independently, but are linked via the fact that the same hyperedges connect to the red, green and blue vertices

Degree sum rules

We have three degree distributions:

- $p_r(k)$ as the fraction of red vertices with degree k,
- $p_g(k)$ the fraction of green vertices with degree k
- $p_b(k)$ the fraction of blue vertices with degree k

These distributions satisfy the sum rules

$$\sum_{k=0}^{\infty} p_r(k) = \sum_{k=0}^{\infty} p_g(k) = \sum_{k=0}^{\infty} p_b(k) = 1,$$

and

$$\sum_{k=0}^{\infty} kp_r(k) = \langle k_r \rangle, \quad \sum_{k=0}^{\infty} kp_g(k) = \langle k_g \rangle, \quad \sum_{k=0}^{\infty} kp_b(k) = \langle k_b \rangle.$$

Hyperedges degree I

We define the *degree of the hyperedges* as the number of neighbours of a given hyperedge

Hyperedges degree

The number *hh* of neighbour of a given hyperedge can be obtained by the previous quantities

$$h \equiv k_r + k_g + k_b - k_c - k_y - k_m$$

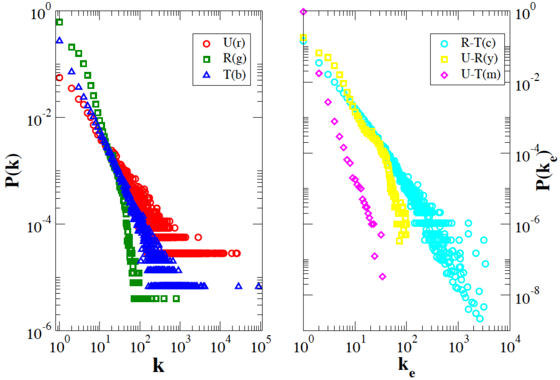
Hhyperedges degree II

Let $P(h)$ represent the fraction of hyperedges connected to exactly hh other hyperedges.
In the absence of correlations between the degrees of the vertices and the edges we have

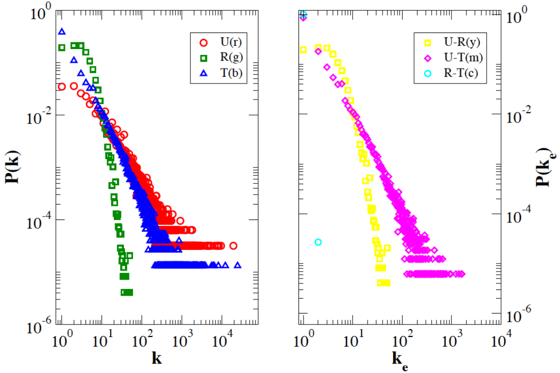
$$
\begin{aligned}
P(hh) \quad = \quad & \sum_{k_r,k_g,k_b,k_c,k_m,k_y} P(k_r)P(k_g)P(k_b)P(k_c)P(k_m)P(k_y) \\
& \cdot\Theta(k_r - k_m - k_y)\Theta(k_g - k_c - k_y) \\
& \cdot\Theta(k_b - k_m - k_c)\delta_{hh,k_r+k_g+k_b-k_c-k_m-k_y}
\end{aligned}
$$

- $\Theta(x)$ is the Heaviside's step function
- $\delta_{x,y}$ is the Kronecker's delta.

Number of hyperedges per vertex/edge in Citeulike

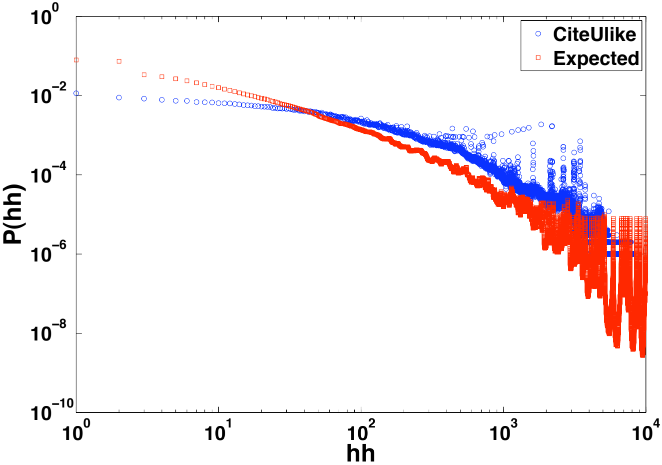


Number of hyperedges per vertex/edge in Flickr

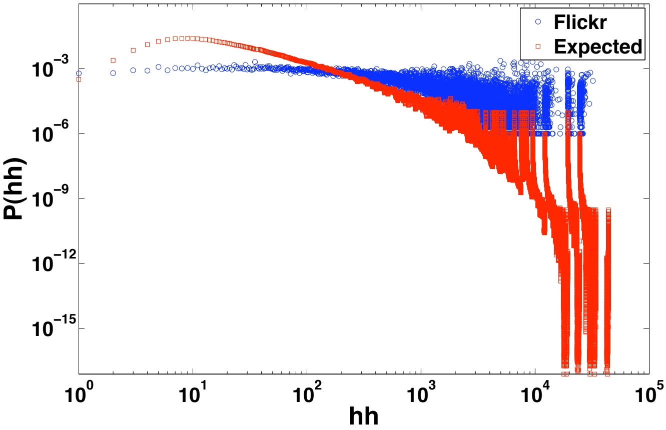


The absence of points for cyan edge (resource-tag) is because tags in Flickr are public and this prevents redundant tagging.

Number of hyperedges per hyperedge in Citeulike



Number of hyperedges per hyperedge in Flickr

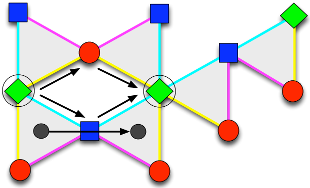


Distance I

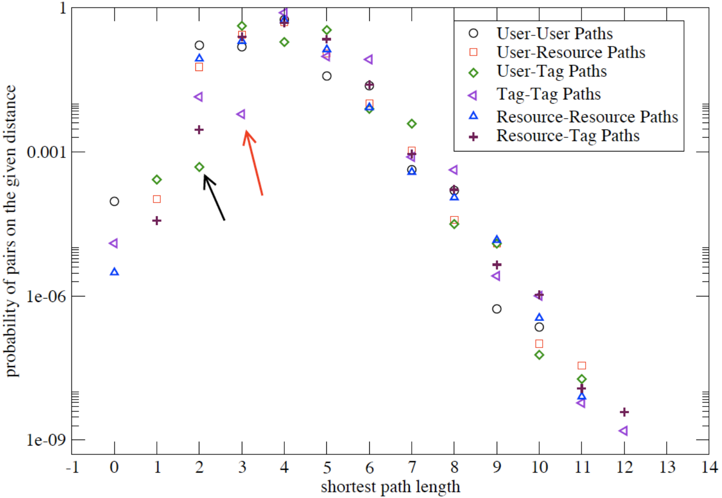
Different possible choices

For vertices/edges

- minimal number of hyperedges which connect vertices/edges
- minimal number of edges between the vertices/edges



Distance II


1
0.001
1e-06
1e-09
probability of pairs on the given distance
-1
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
shortest path length
User-User Paths
User-Resource Paths
User-Tag Paths
Tag-Tag Paths
Resource-Resource Paths
Resource-Tag Paths

Clustering I

In this case again we can use hyperedges to address the connections between vertices.

Coordination number

We introduce the *coordination number* z as the number of immediate neighbors of any color that are connected to it via regular edges

Clustering II

Two immediate bounds can be computed

- Upper bound $z_{max} = 2h$ where h is the number of hyperedges it belongs to
- Lower bound $z_{min} \approx 2\sqrt{h}$

$$z_{min} = \begin{cases} 2n & \text{if} \quad n(n-1) \leq h \leq n^2 \\ 2n+1 & \text{if} \quad n^2 \leq h \leq n(n+1) \end{cases}$$

Clustering III

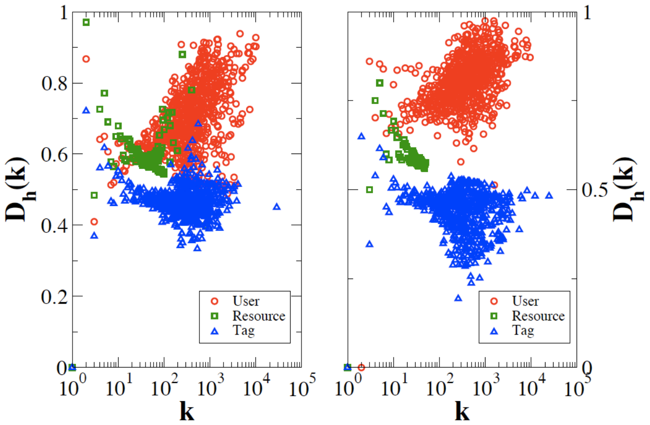
Based on the coordination number defined above for a vertex of degree k, we define a local measure of overlap or clustering,

Hyperedge density

the *hyperedge density* $D_h(k)$:

$$D_h(k) = \frac{z_{max} - z}{z_{max} - z_{min}}.$$

Hyperedges Density in Data



On the left Citeulike network, on the right the Flickr one.

Community structure

Among the various possible methods we clustered together similar vertices

Vertex Similarity

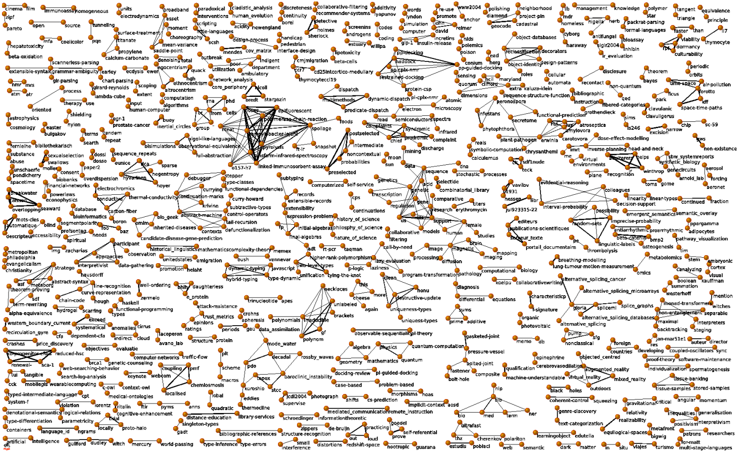
we can define a vertex “distance” as

$$d(v_1, v_2) = \frac{(N_1 \cup N_2) - (N_1 \cap N_2)}{(N_1 \cup N_2) + (N_1 \cap N_2)},$$

where N_1 and N_2 are neighbors of the vertices v_1 and v_2 respectively.

and then connect all the vertices below a certain threshold

Communities II



Hypothesis

- Consider a model hypergraph with n_r red vertices, n_g green vertices, and n_b blue vertices; all with $\langle k_r \rangle, \langle k_g \rangle, \langle k_b \rangle$ mean degree(respectively).
- Each vertex is assigned a degree, corresponding to the number of hyperedges it will have, these degrees can be thought as “stubs”.
- A total of m three-way hyperedges are now created by choosing trios of stubs uniformly at random, one each from a red, green, and blue vertex, and connecting them to form hyperedges.

$$n_r \langle k_r \rangle = n_g \langle k_g \rangle = n_b \langle k_b \rangle = M.$$

Expected values

Given that there are m hyperedges in total, the overall probability of a hyperedge between i, j, and k is then

$$P_{ijk} = M \times \frac{k_i}{M} \times \frac{k_j}{M} \times \frac{k_k}{M} = \frac{k_i k_j k_k}{M^2}.$$

Via a similar argument, the probability that there is a hyperedge connecting a particular red/green pair i, j (or any other color combination) is $\frac{k_i k_j}{M}$

Excess degree distribution

We are interested in the probability that by following an hyperedge you end up in a vertex involved in other k hyperedges other than the one we followed.
(i.e. Excess degree= degree-1)

$$q_r(k) = \frac{(k+1)p_r(k+1)}{\langle k_r \rangle}$$

Generating Functions I

We begin by defining generating functions for the degree distributions

$$r_0(z) = \sum_{k=0}^{\infty} p_r(k) z^k$$

We now define the *generating functions* for the excess degree distributions:

$$r_1(z) = \sum_{k=0}^{\infty} q_r(k) z^k = \frac{1}{\langle k_r \rangle} \sum_{k=0}^{\infty} (k+1) p_r(k+1) z^k = \frac{r_0'(z)}{r_0'(1)}$$

and similarly for b and g

Generating Functions II

Projections

The Generating Functions can be used to compute the degree distribution on projected graphs.

I.e. take a red vertex A

- it has s green neighbours (s distributed as $p_r(s)$)
- any of the s has t_s red neighbours (apart from A and t following $q_g(t)$).

the probability that A has k neighbours in the projection is

$$\rho_g(k) = \sum_{s=0}^{\infty} p_r(s) \sum_{t_1=0}^{\infty} q_g(t_1) .... \sum_{t_s=0}^{\infty} q_g(t_s) \delta\left(k, \sum_{s=0}^{\infty} p_r(s)\right)$$

Projections II

Multiplying by z^k and summing over k we have

$$R_g(z) = \sum_{k=0}^{\infty} z^k \rho_g(k)$$

that becomes

$$R_g(z) = r_0[g_1(z)]$$

Projections II

Multiplying by z^k and summing over k we have

$$R_g(z) = \sum_{k=0}^{\infty} z^k \rho_g(k)$$

that becomes

$$R_g(z) = r_0[g_1(z)]$$

$$\begin{aligned}
R_g(z) &= \textstyle\sum_{k=0}^{\infty} z^k \sum_{s=0}^{\infty} p_r(s) \sum_{t_1=0}^{\infty} q_g(t_1) \dots \sum_{t_s=0}^{\infty} q_g(t_s)\, \delta\left(k, \sum_{n=1}^{s} t_n\right) \\
&= \textstyle\sum_{s=0}^{\infty} p_r(s) \sum_{t_1=0}^{\infty} q_g(t_1) \dots \sum_{t_1=0}^{\infty} q_g(t_s) z^{\sum_{n=1}^{s} t_n} \\
&= \textstyle\sum_{s=0}^{\infty} p_r(s) \sum_{t_1=0}^{\infty} q_g(t_1) z_1^t \dots \sum_{t_1=0}^{\infty} q_g(t_s) z^{t_s} \\
&= \textstyle\sum_{s=0}^{\infty} p_r(s) \left[\sum_{t=0}^{\infty} q_g(t) z^t\right]^s \\
&= r_0[g_1(z)]
\end{aligned}$$

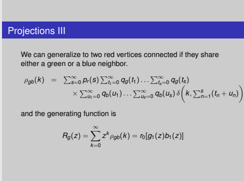
Projections III

We can generalize to two red vertices connected if they share either a green or a blue neighbor.

$$\rho_{gb}(k) = \sum_{s=0}^{\infty} p_r(s) \sum_{t_1=0}^{\infty} q_g(t_1) \ldots \sum_{t_s=0}^{\infty} q_g(t_s) \times \sum_{u_1=0}^{\infty} q_b(u_1) \ldots \sum_{u_s=0}^{\infty} q_b(u_s)\, \delta\left(k, \sum_{n=1}^{s}(t_n+u_n)\right)$$

and the generating function is

$$R_g(z) = \sum_{k=0}^{\infty} z^k \rho_{gb}(k) = r_0[g_1(z)b_1(z)]$$



$$
\begin{aligned}
R_{gb}(z) &= \sum_{k=0}^{\infty} z^k \sum_{s=0}^{\infty} p_r(s) \sum_{t_1=0}^{\infty} q_g(t_1) \ldots \sum_{t_s=0}^{\infty} q_g(t_s) \\
&\qquad \times \sum_{u_1=0}^{\infty} q_b(u_1) \ldots \sum_{u_s=0}^{\infty} q_b(u_s)\, \delta\left(k, \sum_{n=1}^{s}(t_n+u_n)\right) \\
&= \sum_{s=0}^{\infty} p_r(s) \left[\sum_{t=0}^{\infty} q_g(t) z^t\right]^s \left[\sum_{u=0}^{\infty} q_b(u) z^u\right]^s \\
&= r_0(g_1(z) b_1(z)).
\end{aligned}
$$

Scale-free Graphs

We use generating function to compute the Degree distribution in particular

Degree from Generating Functions

$$p_k = \frac{1}{k!} \frac{d^k R_{gb}}{dz^k}_{|z=0}$$

Random Graph

Consider a tripartite random graph with Poisson degree distributions thus:

$$p_r(k) = e^{-\langle k_r \rangle} \frac{\langle k_r \rangle^k}{k!}, \quad p_g(k) = e^{-\langle k_g \rangle} \frac{\langle k_g \rangle^k}{k!}, \quad p_b(k) = e^{-\langle k_b \rangle} \frac{\langle k_b \rangle^k}{k!},$$

The generating function for this distribution is given by

$$R_{gb} = r_0(g_1(z)b_1(z)) = e^{\langle k_r \rangle (e^{(\langle k_g \rangle + \langle k_b \rangle)(z-1)} - 1)}.$$

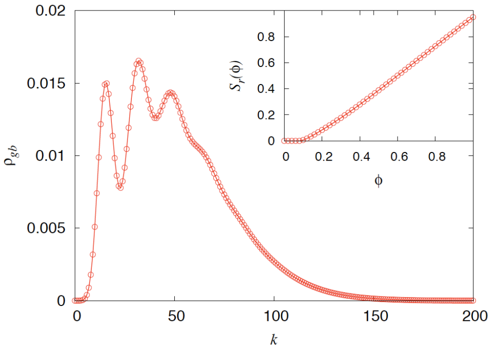
Random Graph

Expanding in powers of z, we then find that the probability $\rho_{gb}(k)$ of a red vertex having exactly k neighbors in the projected network is

$$\rho_{gb}(k) = \frac{(\langle k_g\rangle+\langle k_b\rangle)^k}{k!} e^{\langle k_r\rangle(e^{-(\langle k_g\rangle+\langle k_b\rangle)}-1)} \times \sum_{m=1}^{k} \left\{ {k \atop m} \right\} \left[\langle k_r\rangle e^{-(\langle k_g\rangle+\langle k_b\rangle)}\right]^m,$$

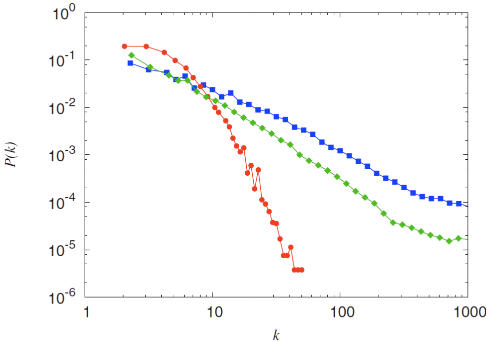
where $\left\{ {k \atop m} \right\}$ is a Stirling number of the second kind, i.e., the number of ways of dividing k objects into m nonempty sets

Random Graph Results



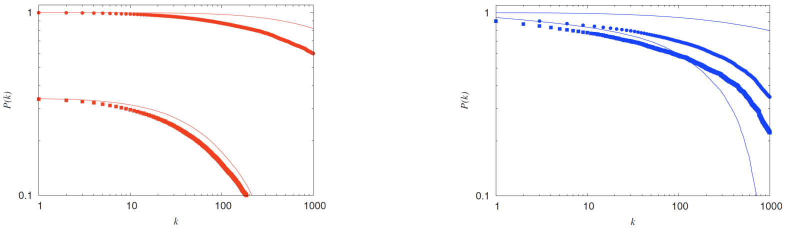
The degree distribution for the projection of the Poisson hypergraph onto its red vertices alone.

Scale-free Graphs Data

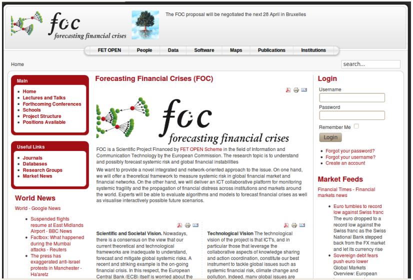


Experimentally the distributions are power-law

Scale-free Graphs Results



Positions



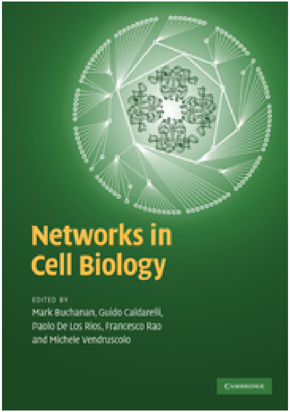
http://www.focproject.net

Financial Networks

Try to forecast avalanches and decide who's to bail out

- CNR (Rome),
- U. Marche (Ancona, I),
- ETH (Zürich, CH),
- CITY (London Uk),
- Said Business School (Oxford UK),
- FBM (Barcelona, SPAIN),
- ECB (Frankfurt, EU)

Advertisement nr. 2



Networks in Cell Biology

Edited by Mark Buchanan, Guido Caldarelli, Paolo De Los Rios, Francesco Rao, Michele Vendruscolo

Summary

- We can describe tagged networks as hypergraphs, that is graphs where an hyperedge connects more than one vertex.
- This natural description allows to detect deviation from random hypergraph model used as a reference null case.
- We find correlations between vertices not described by the simple degree distributions.

- Outlook
  - Generalize the approach to interacting networks not composed by regular triples
  - Explore the fragility issues based on hyperedges analysis

For Further Reading I

- G. Caldarelli.
  Scale-Free Networks, Oxford University Press, 2007.
- R. Lambiotte, M. Ausloos, *Lecture Notes in Computer Science*, **3993**, 1114, (2006).
- C. Cattuto, V. Loreto, L. Pietronero, *Proc. Nat. Acad. Sci. U.S.A.* **104**, 1461, (2007).
- O. Görlitz, S. Sizov S. Staab, *Lecture Notes in Computer Sciences* **5021**, 807-811, (2008).
- G. Ghoshal, V. Zlatić, G. Caldarelli and M.E.J. Newman, *Phys. Rev. E*, **79**, 066118, (2009).