Random Hypergraphs and their applications
Study of topology in tagged social networks

G. Ghoshal\textsuperscript{1} M.E.J. Newman\textsuperscript{1} V. Zlatić\textsuperscript{2}
\textbf{G. Caldarelli}\textsuperscript{2,3}

\textsuperscript{1}Department of Physics and Michigan Center for Theoretical Physics,
University of Michigan, Ann Arbor Michigan USA

\textsuperscript{2}CNR Istituto dei Sistemi Complessi, University “Sapienza” Rome Italy

\textsuperscript{3}London Institute for Mathematical Sciences, London Mayfair (visiting)

CABDyN Oxford January 18\textsuperscript{th} 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

---
²G. Ghoshal, V. Zlatić, GC *PRE* 79 066118 (2009)
Tagged Networks

Delicious site

The tastiest bookmarks on the web.
Save your own or see what's fresh now!

Search the biggest collection of bookmarks in the universe...

Explore Tags

Tags are keywords people use to describe their bookmarks

Guido Caldarelli
CNR Institute of Complex Systems, Dep. Physics Sapienza University, Rome Italy

Hypergraphs and tagged social networks
Citeulike site

citeulike is a free service for managing and discovering scholarly references

3,684,935 articles - 3,390 added today.

- Easily store references you find online
- Discover new articles and resources
- Automated article recommendations NEW
- Share references with your peers
- Find out who's reading what you're reading
- Store and search your PDFs

Join now
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 a generalization of graphs,
- Hyperedges are arbitrary set of vertices,
- Tagged systems are 3-uniform hypergraphs.
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
One approach is “to project” this triple structure along one of the three components. Similarly to bipartite graphs of collaborations for ordinary graphs.
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
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:

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

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.
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

$$\sum_{k=0}^{\infty} kp_r(k) = \langle k_r \rangle, \quad \sum_{k=0}^{\infty} kp_g(k) = \langle k_g \rangle, \quad \sum_{k=0}^{\infty} kp_b(k) = \langle k_b \rangle.$$
We define the *degree of the hyperedges* as the number of neighbours of a given hyperedge.

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

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

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

- $\Theta(x)$ is the Heaviside’s step function
- $\delta_{x,y}$ is the Kronecker’s delta.
Number of hyperedges per vertex/edge in Citeulike

Guido Caldarelli
CNR Institute of Complex Systems, Dep. Physics Sapienza University, Rome Italy

Hypergraphs and tagged social networks
Hyperedges

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.
Hyperedges

Number of hyperedges per hyperedge in Citeulike

Guido Caldarelli
CNR Institute of Complex Systems, Dep. Physics Sapienza University, Rome Italy

Hypergraphs and tagged social networks
Number of hyperedges per hyperedge in Flickr

![Graph showing the distribution of hyperedges per hyperedge in Flickr. The x-axis represents the number of hyperedges (hh), while the y-axis represents the probability P(hh). The graph compares the actual data from Flickr with an expected distribution.](image)
Distance

Different possible choices
For vertices/edges
- minimal number of hyperedges which connect vertices/edges
- minimal number of edges between the vertices/edges
**Distance II**

The image shows a graph with the x-axis representing the shortest path length and the y-axis representing the probability of pairs on the given distance. The graph plots various types of paths, such as User-User Paths, User-Resource Paths, User-Tag Paths, Tag-Tag Paths, Resource-Resource Paths, and Resource-Tag Paths, using different markers and colors.

Guido Caldarelli

CNR Institute of Complex Systems, Dep. Physics Sapienza University, Rome Italy

Hypergraphs and tagged social networks
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 $z_{min} \approx 2\sqrt{h}$

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

Hypergraphs and tagged social networks

Guido Caldarelli
CNR Institute of Complex Systems, Dep. Physics Sapienza University, Rome Italy
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)

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

$$\rho_g(k) = \sum_{s=0}^{\infty} p_r(s) \sum_{t_1=0}^{\infty} q_g(t_1) \cdots \sum_{t_s=0}^{\infty} q_g(t_s) \delta \left( k, \sum_{s=0}^{\infty} p_r(s) \right)$$
Multiplying by $z^k$ and summing over $k$ we have

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

that becomes

$$R_g(z) = r_0[g_1(z)]$$
Multiplying by $z^k$ and summing over $k$ we have

$$R_g(z) = \sum_{k=0}^{\infty} z^k \sum_{s=0}^\infty p_r(s) \sum_{t_1=0}^{\infty} q_g(t_1) \cdots \sum_{t_s=0}^{\infty} q_g(t_s) \delta\left(k, \sum_{n=1}^{s} t_n\right)$$

that becomes

$$R_g(z) = r_0[g_1(z)]$$
We can generalize to two red vertices connected if they share either a green or a blue neighbor.

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

and the generating function is

$$
R_g(z) = \sum_{k=0}^{\infty} z^k \rho_{gb}(k) = r_0[g_1(z)b_1(z)]
$$
We can generalize to two red vertices connected if they share either a green or a blue neighbor.

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

and the generating function is

\[ R_{gb}(z) = \sum_{k=0}^{\infty} z^k \sum_{s=0}^{\infty} p_r(s) \sum_{t_1=0}^{\infty} q_g(t_1) \cdots \sum_{t_s=0}^{\infty} q_g(t_s) \times \sum_{u_1=0}^{\infty} q_b(u_1) \cdots \sum_{u_s=0}^{\infty} q_b(u_s) \delta \left( k, \sum_{n=1}^{s} (t_n + u_n) \right) \]

\[ = \sum_{s=0}^{\infty} p_r(s) \left[ \sum_{t=0}^{\infty} q_g(t) z^t \right]^s \left[ \sum_{u=0}^{\infty} q_b(u) z^u \right]^s \]

\[ = r_0(g_1(z)b_1(z)). \]
We use generating function to compute the Degree distribution in particular

Degree from Generating Functions

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

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

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

where $\left\{ \begin{array}{c} k \\ m \end{array} \right\}$ is a Stirling number of the second kind, i.e., the number of ways of dividing $k$ objects into $m$ nonempty sets.
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
Comparison with real data

Scale-free Graphs Results

Guido Caldarelli
CNR Institute of Complex Systems, Dep. Physics Sapienza University, Rome Italy

Hypergraphs and tagged social networks
Financial Networks

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

Guido Caldarelli

CNR Institute of Complex Systems, Dep. Physics Sapienza University, Rome Italy

Hypergraphs and tagged social networks
Advertisement nr. 2

Networks in Cell Biology

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

Guido Caldarelli
CNR Institute of Complex Systems, Dep. Physics Sapienza University, Rome Italy

Hypergraphs and tagged social networks
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

G. Caldarelli. 


