WOLFRAM SCIENCE
SUMMER SCHOOL 2013

What Popularity tells us about a Wikipedia Sub-Graph?

Valentina Biagini

University “La Sapienza” of Rome

Project description

Studying the popularity time series of two Wikipedia pages and their correlation, can we say if this two pages are linked? And how strong is their link?

General Characteristics of Popularity Time Series

Looking at the time series we observed a global behaviour, we studied an independent set of 10000 pages in a time window of two months. We saw a ***weekly effect***, there is weekly fluctuation of the number of visitors, about 20-30%. We calculated 7 daily coefficient in order to correct Popularity and reduce the overestimation of correlations. Looking at an independent set of 1000 pages in a time window of a year, we observed a *seasonal effect*, but in our time window this is negligible.

What happens when there is an Event? Does Popularity diffuse?

I chose a recent event with a peak around the release date of the movie “Star Trek Into Darkness” (May 16th 2013). I built the weighted sub-graph of Wikipedia English pages, starting from the page of the movie. I collected all the time series of this pages in a time window of two month, around the Event. Looking at the time series we observed that, at this daily resolution, there is ***no propagation of Popularity***, there is no delay between two time series. With this data we can’t consider to study a dynamic. And also we can’t study causal relations. For this reason we decided to study the relation between different pages Popularity with the Pearson Correlation.

Characteristics of The Wikipedia Sub-Graph

The graph in analysis is extracted starting from a central page and considering only English pages. We considered all the out-links until the second neighbors, and we selected all the links with at least 2 repetitions. Multiple links are very common and we used this as a measure of the strenght of the connection, the ***Links Multiplicity***. We obtained a weighted directed Graph of 3126 Edges and 2164 Vertices.

Can we predict the real connections of the Graph from Popularity?

The correlations between two time series imply a symmetric Graph:

N_Links (v , u) ~ 1 / (1 - Correlation (v , u))

We observed an interesting relation between the total number of links of a page and its average Popularity:

N_Links(v) ~ Popularity (v)^0.5

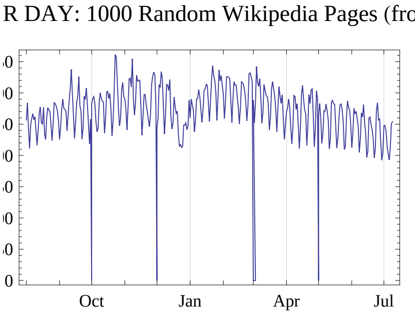
Correlations will give us symmetric information and Popularity asymmetric information on the Graph. *Our hypotesis is a prediction rule of this form:*

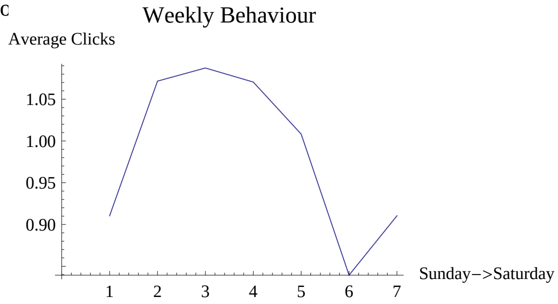
Multiplicity (u , v) ~ [1 / (1 - Correlation (u , v)]^a [Popularity (u)]^b [Popularity (v)]^c

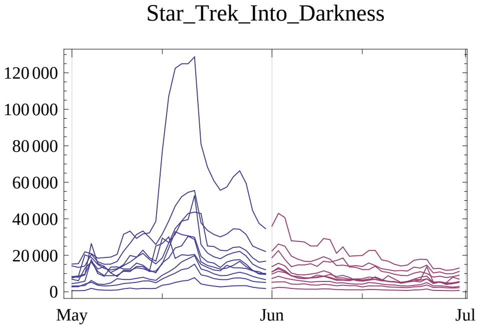
We tested our prediction rules with different coefficients and we defined an error measure.

Graphics

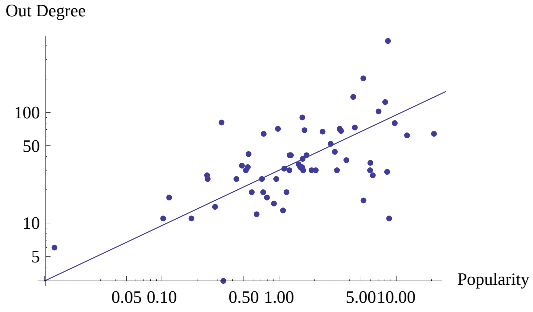
»

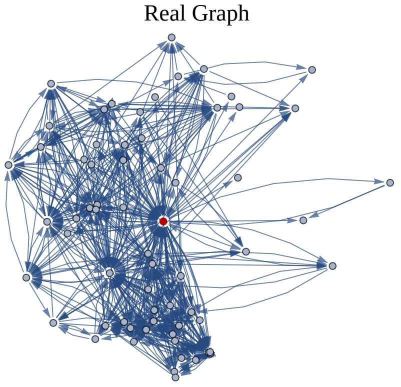




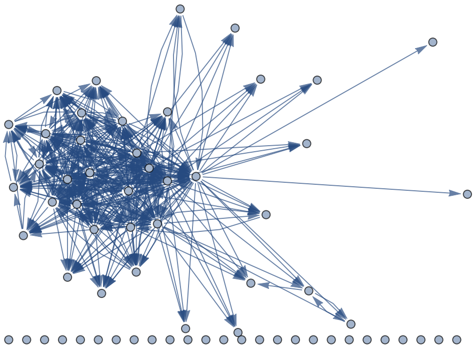


N_Links(v) ~ Popularity (v)^0.5





Predicted Graph



Summary of results and conclusions

We observed a global behaviour in the Popularity Time Series, we mesured thi effect with 7 daily coefficients that we used to reduce the overestimation of the correlations. We observed that at this daily resolution the popularity due to a big event that involve one page doesn’t diffuse over the connected pages. We can’t study the dynamic and we can’t study the causal relations. We can study

correlations, but we can't say if the correlation between two pages exists because of the structure of the network, or because there is a causal connection between the two pages, beyond the network. This two mechanisms probably cohexist but with this resolution we can't see if one is predomninant. We decided to study the Pearson Correlation.
We built a Wikipedia Sub-Graph starting from a central page and we used Links Multipilicty as a mesure of the strenght of the connection. We cosidered the time series of each page in our Sub-Graph, in a time window of two months, and we calculated correlatons and average Popularity. We observed that the number of links of a page is related to its popularity by a square root law. Correlations give us a symmetric Graph, and to improve our prediction we used the average Popularity, that gives us asymmetric information.
We defined a set of Prediction Rule and an error mesure and we find a satisfiyng prediction with this coefficients and a threshold of at least 1. link predicted:

$$\mathit{Multiplicity}\,(u\,,v\,) = \frac{0.5}{[(1-\mathit{Correlation}\,(u\,,v\,)\,]}\,[\,\mathit{Popularity}\,(\,u\,)\,]^{0.5}\,[\,\mathit{Popularity}\,(\,v\,)\,]^{0.1}$$

this gives us a typical error of ± 1.45 links and a mean difference between predicted links and real links of -0.12.

Future directions

It would be interesting study the dynamic, starting from a higher resolution of the Time Series, for instance a resolution of minutes, and see if there is a diffusion. And also try to understand the causal relations that are hidden in this system, starting from the dynamic. It would be interesting also have the history of the Wikipedia structure, in order to better understand the causal effect between correlations and connections, and use this hystorical data to test and improve our prediction rules. An interesting future application could be to suggest a website like Wikipedia to connect pages that aren't connected jet, but that are connected on the basis of our prediction.