What Popularity tells us about a Wikipedia Sub-Graph?

About this work

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

- **Characteristics of The Wikipedia Sub-Graph**
  The graph in analysis is extracted starting from a central page and considering only English pages. We considered a recent event that involves the central page. 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 strength of the connection, the ***Links Multiplicity***. We obtained a weighted directed Graph of 3126 Edges and 2164 Vertices.
- **General Characteristics of Popularity Time Series**
  Looking at the time series we observed a global behavior, 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.
- **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
  Our hypothesis is a prediction rule of this form:
  $$Multiplicity(u,v) \sim [1/(1-Correlation(u,v)]^a\ [Popularity(u)]^b\ [Popularity(v)]^c$$
  We tested our prediction rules with different coefficients and we defined an error measure.

The Wikigraph

Useful Functions

Starting from an initial Wikipedia Page, this group of functions built the weighted graph of depth 2 (we consider only links with multiplicity > 1). Multiplicity is our weight.

Operations on the Graph

Wikigraph of our Data Set

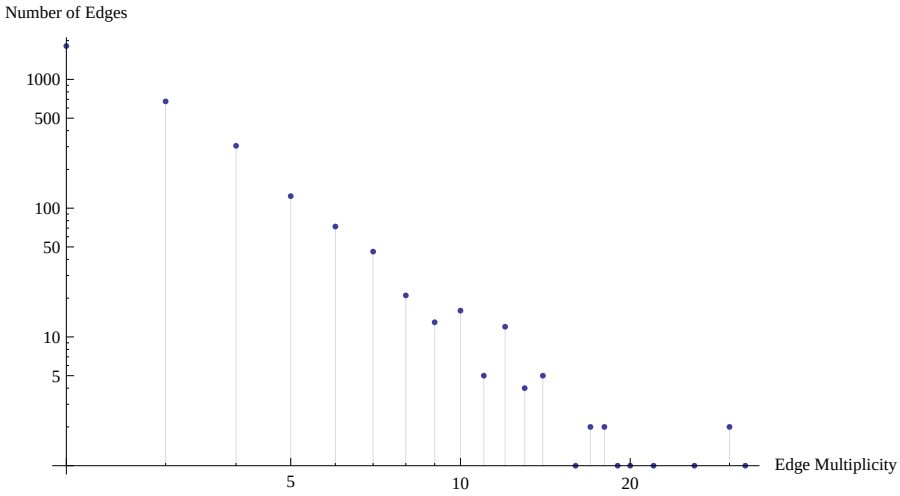
Star Trek Wikigraph: 3126 Edges, 2164 Vertices

We select the most relevant links in each page, starting from "Star Trek Into Darkness", and we take pages until its second neighbors. We consider multiplicity of links as a measure of their importance. We filter pages using Multiplicity > 1.

```
StarTrekNet = NewGraphWithCorrectMultiplicity[
    Import["/Users/Levantina/Documents/WOLFRAM/PROJECT/startrekNetwork/
       allStarTrekWeighted.tsv", "TSV"]];
```

Edge Multiplicity

```
ListLogLogPlot[Tally[StarTrekNet[[All, 3]]], Filling → Axis,
 PlotRange → All, AxesLabel → {"Edge Multiplicity", "Number of Edges"}]
```



I consider the unweighted graph:

```
UnweightedST = StarTrekNet[[All, ;; 2]];
```

```
STGraph = Graph[DirectedEdge @@@ UnweightedST];
```

```
STVertices = VertexList[STGraph];
```

K-Cores decomposition

Popularity Time Series

Useful Functions

What about the Popularity behavior?

Popularity Time Series of our Data Set

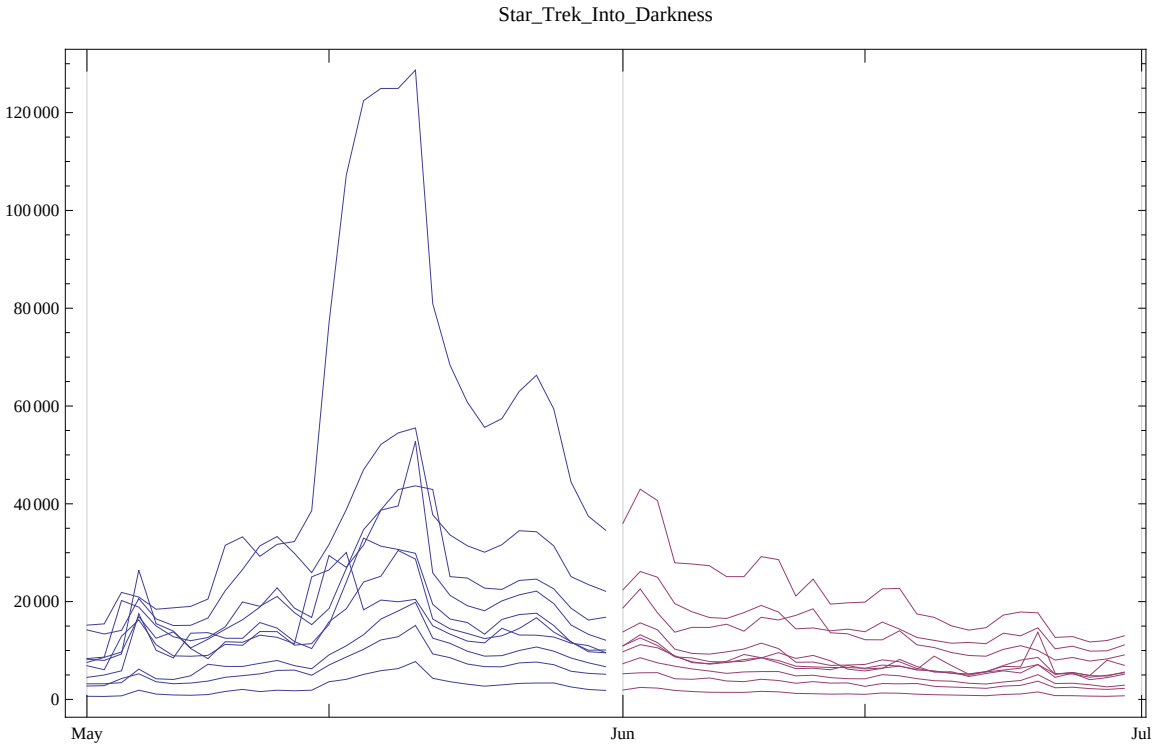
Import Data

Interesting Event

Time Series of the others vertices in the Wikigraph

Given an event, are Popularities Delayed in time?

Popularities for neighbors of “Star Trek Into Darkness”, “”Star Trek (film)”, “Star Trek”, “Spock”, ““James T. Kirk”. They don’t seem delayed in time. For this reason to study correlations we will use the Pearson Correlation.



Instant Behavior

Our resolution is of one day, we can’t see a diffusion in this behavior, we can’t consider a dynamic. And also we can’t study causal relations. We can see 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 mechanism probably coexist but we can’t see with these data if one is predominant.

Correlations between Time Series

Useful Functions

Fast Correlation

Adjacency Matrix

Evaluation with the daily correction

Daily correction for Time Series

We noticed a weekly fluctuation of the visitors on Wikipedia, we found a weekly periodicity. We have a correction for each day of the week. And during the evaluation of the correlations we will consider this 7 coefficients. In order to reduce the overestimating of the correlations.

```
averagesDaily = {1.0861167208459364`,
  1.070509192186192`, 1.0082207459429013`, 0.8393580386204945`,
  0.910410111634681`, 1.0716210490409843`, 1.0873540902600047`}
```

```
{1.08612, 1.07051, 1.00822, 0.839358, 0.91041, 1.07162, 1.08735}
```

```
rightSTPop = Flatten[cleanedSTTimeSeries[[#, ;; 2, All, 2]]] & /@
   Range[1, Length[cleanedSTTimeSeries[[All, ;; 2, All, 2]]]];
```

```
Corrections[daily_,timeseries_]:=
Module[{days=Length[timeseries[[1]]]},
    Map[#/Flatten[Table[daily,{Round[days/7]+1}]][[;;Length[#]]] &, timeseries,{1}]
]
```

```
PopularityNormal = rightSTPop;
```

```
PopularityCorrected = Corrections[averagesDaily, rightSTPop];
```

Pearson Correlation with Popularity and Corrected Popularity

Confront Correlations

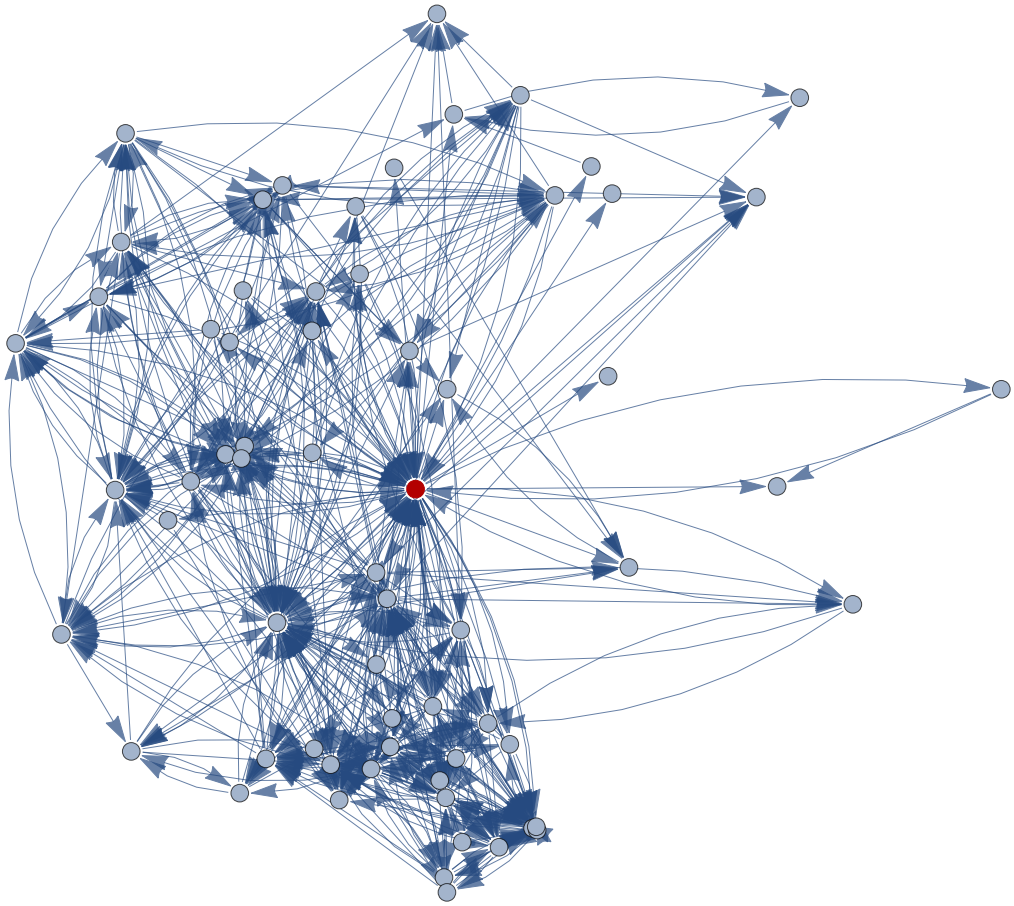
How can we reproduce the real connections in the Graph?

The Neighborhood Graph of ‘Star Trek Into Darkness’: first neighbors.

The Real Sub-Graph

This is what we want to reproduce.

Star Trek Into Darkness



Data

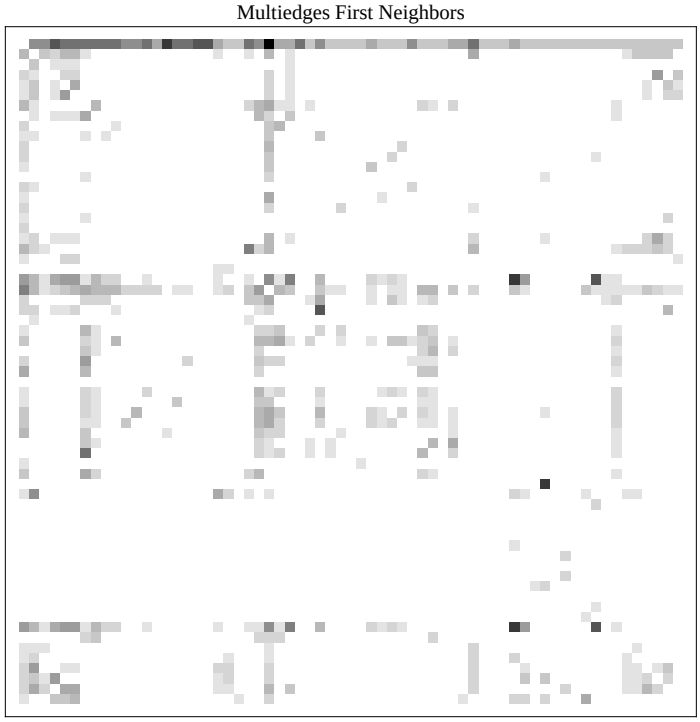
The Real Adjacency Matrix

We build the adjacency matrix of the graph (multiplicity of links will be a measure of their strength):

```
distanceMatrixCorrected =
  Import["/Users/Levantina/Documents/WOLFRAM/PROJECT/startrekNetwork/
     distanceMaxtrixCorrected.tsv", "TSV"];

distanceMatrix65 = distanceMatrixCorrected[[ ;; 65, ;; 65]];
```

This matrix represent the links that we want to reproduce.

Multiedges First Neighbors



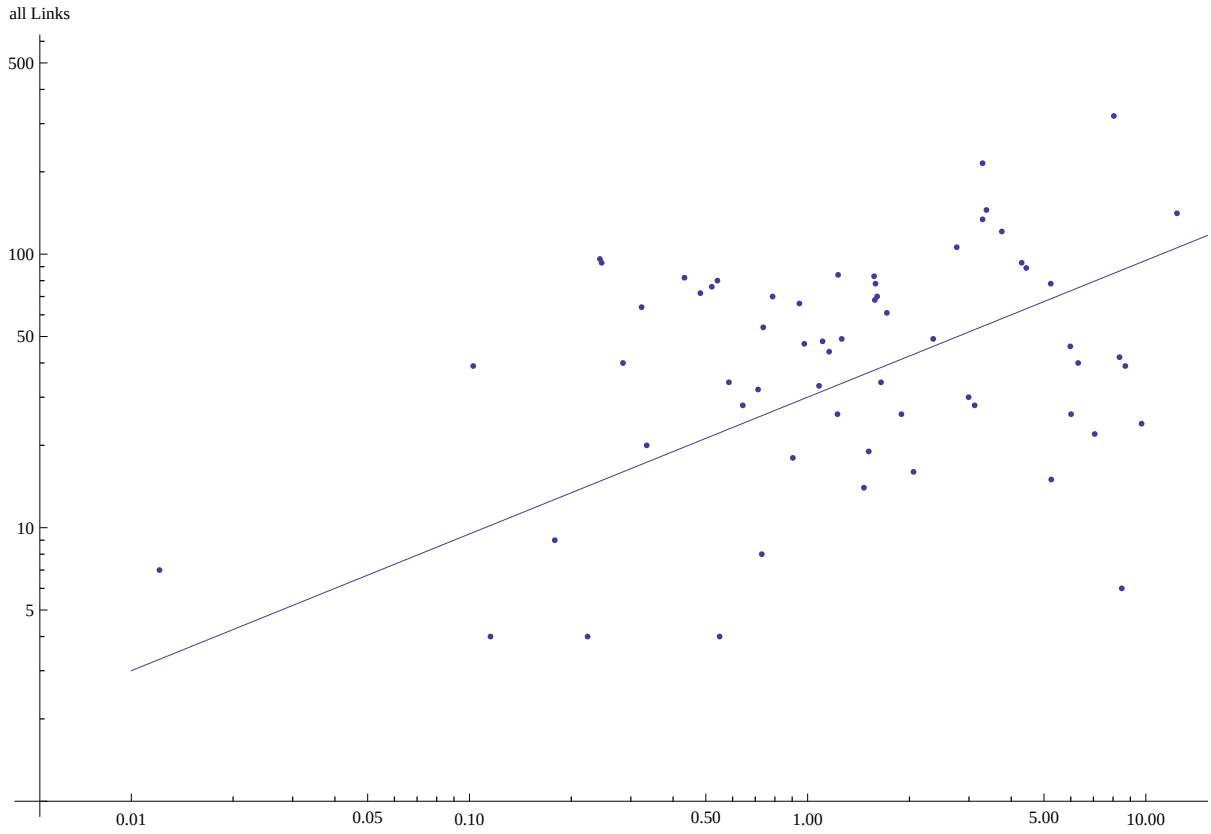
Total Number of Real Links VS Popularity

```
AdjMatr65 = Import[
    "/Users/Levantina/Documents/WOLFRAM/PROJECT/startrekNetwork/adjMatrix65.tsv"
    , "TSV"];
```

```
AdjMatr65 = Partition[Flatten[AdjMatr65], 65]; allLinks =
  (Plus @@ AdjMatr65[[All, #]] + Plus @@ AdjMatr65[[#, All]] ) & /@ Range[1, 65];
Max[allLinks]
```

```
meansSTPopNorm = Flatten@Import[
     "/Users/Levantina/Documents/WOLFRAM/PROJECT/Timeseries/meansSTPopNorm.tsv"
     , "TSV"];
```

I find an interesting behavior: Links(v) ~ Popularity(v)^0.5

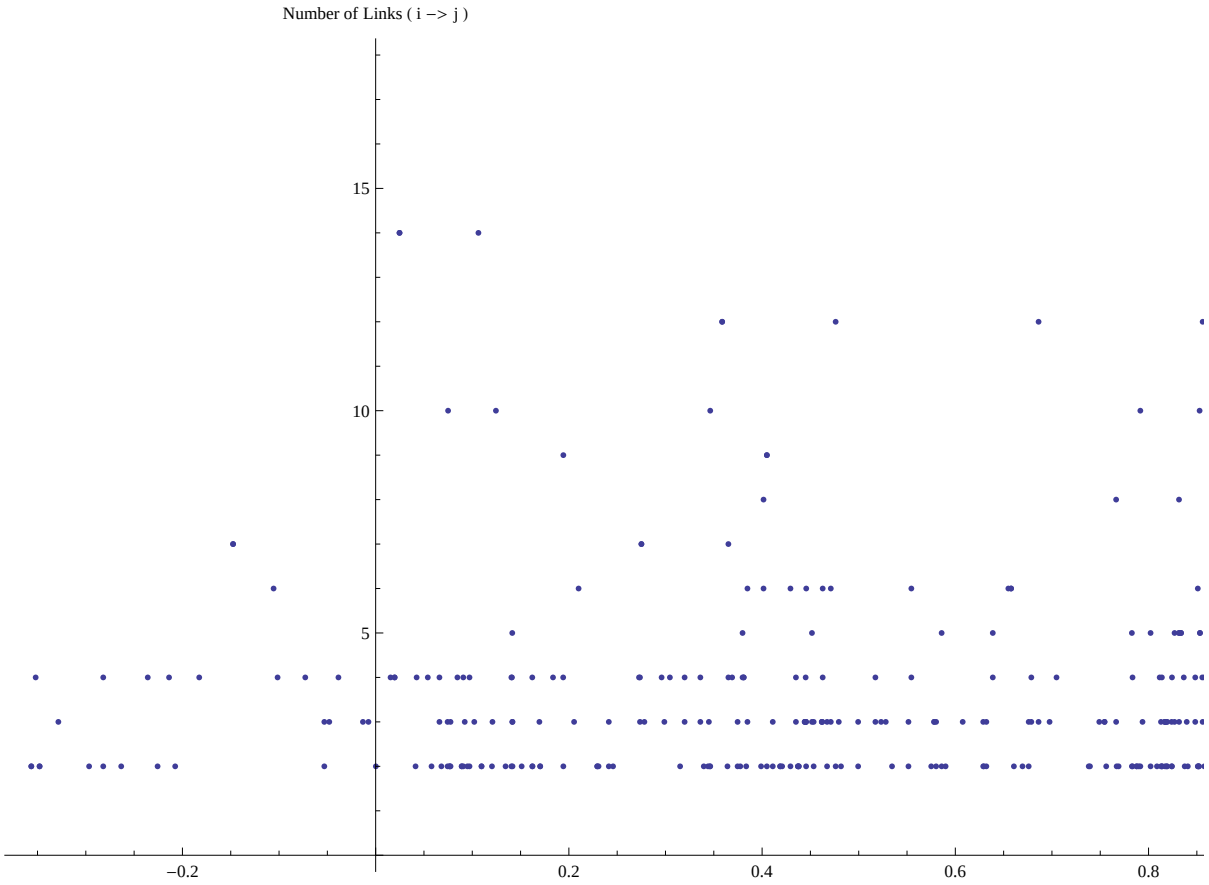


We are going to consider this behavior to predict the number of links that connect two pages.

Number of Real Links VS Correlation

```
corrMatrix = Import["/Users/Levantina/Documents/WOLFRAM/PROJECT/Timeseries/
     CorrelationsSTCorrectedTimeSeries.tsv", "TSV"];

corrMatrixNoSelfCorr =
  corrMatrix * (-1) * (IdentityMatrix[Length[First[corrMatrix]]] - 1);

corrMatrix65 = corrMatrixNoSelfCorr[[ ;; 65, ;; 65]];

corrMatrix65 = Import[
   "/Users/Levantina/Documents/WOLFRAM/PROJECT/Timeseries/corrMatrix65.tsv",
   "TSV"];

meansSTPopNorm = Import[
   "/Users/Levantina/Documents/WOLFRAM/PROJECT/Timeseries/meanSTPopNorm2Months.
     tsv", "TSV"];

Show[Histogram[Cases[
   Transpose[{Flatten[Round[100 * corrMatrix65] / 100.], Flatten[AdjMatr65]}],
   Except[{0., 0}]], PlotRange → All], LogLogPlot[1.5 + 5. x^0.7, {x, 0.001, 2}]]
```

```
$Aborted
```



From Correlations to a Graph

All Correlations Matrix

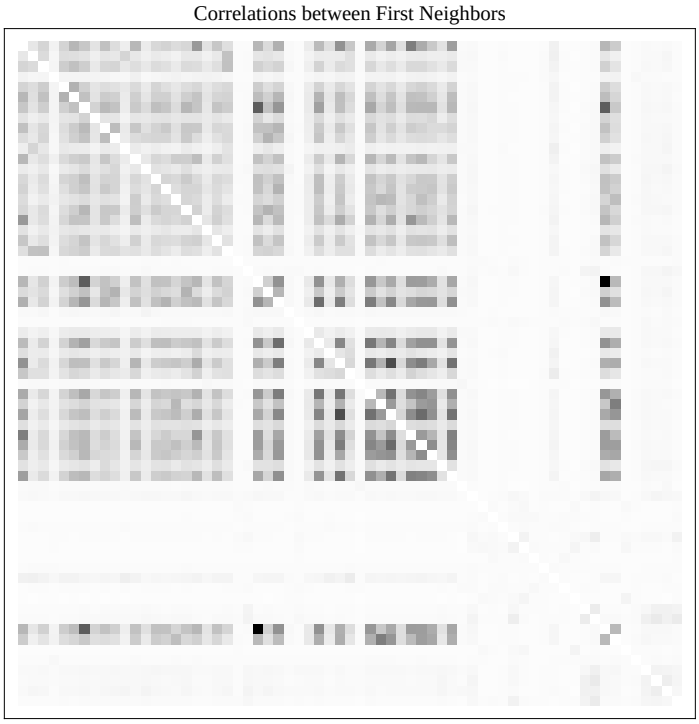
We build the adjacency matrix with all the distances d between the vertices (a measure of the correlation):

d = (1 - corr [i , j])
corr [i , j] = 1 - d

```
distanceMatrixCorrected =
  Import["/Users/Levantina/Documents/WOLFRAM/PROJECT/startrekNetwork/
     distanceMaxtrixCorrected.tsv", "TSV"];

distanceMatrix65 = distanceMatrixCorrected[[ ;; 65, ;; 65]];
```

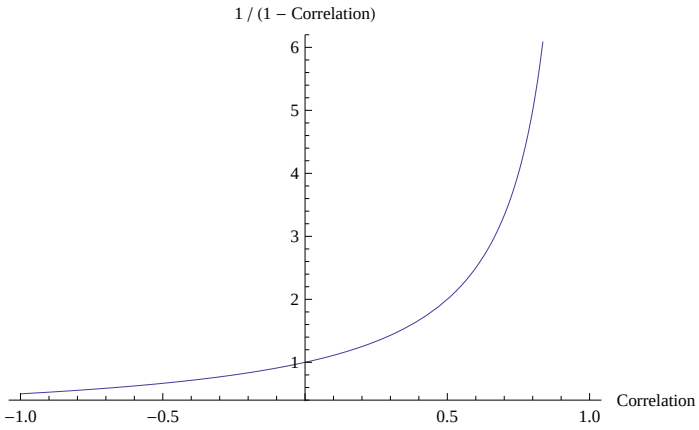
How the “correlation” matrix look like? We consider:
Multiplicity (i , j) ~ 1 / (1 - corr [i , j])

```
ArrayPlot[1. / distanceMatrix65 * removingSelfLoops,
 PlotLabel -> "Correlations between First Neighbors"]
```



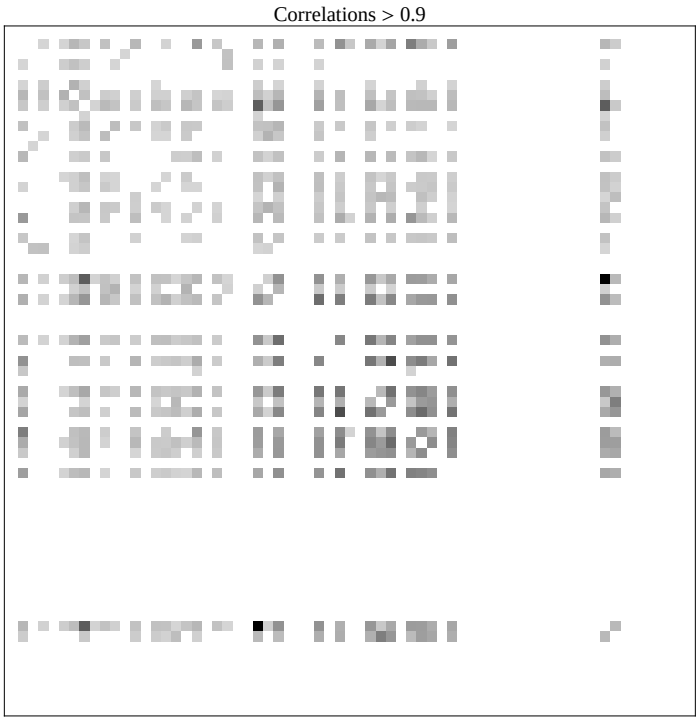
Correlation Matrix with a Threshold

```
Plot[1. / (1 - x), {x, -1., 1},
 AxesLabel → {"Correlation", "1 / (1 - Correlation)"}]
```




```
thresholdInverseDistanceMatrix65 =
  Map[If[# < 10, 0, #] &, (1. / distanceMatrix65) * removingSelfLoops, {2}];
```

```
ArrayPlot[thresholdInverseDistanceMatrix65, PlotLabel → "Correlations > 0.9"]
```

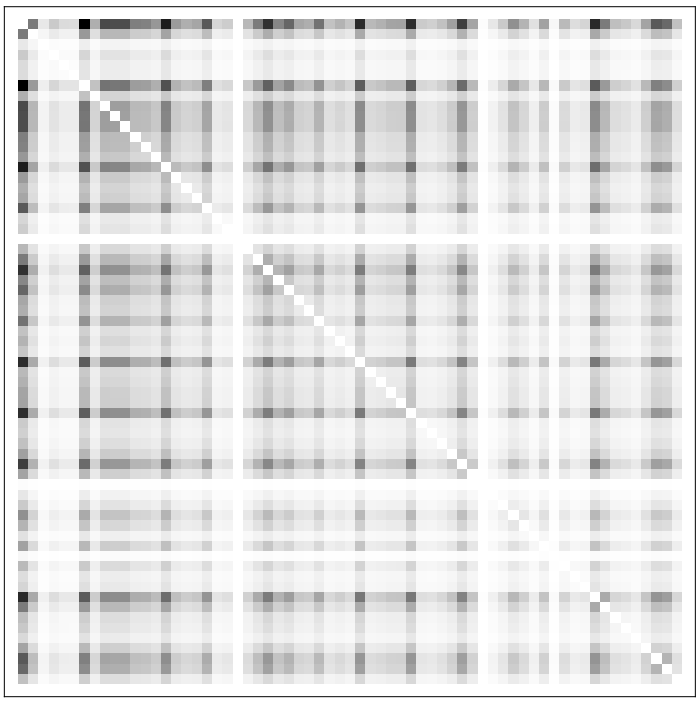


Matrix of the Square Root of the Popularity Product between each Page

How the popularity matrix look like? If we consider:
Multiplicity (i , j) ~ [Pop (i) * Pop (j)] ^ 0.5

```
PopMatrixSqrt =
  Outer[Sqrt[#1 #2] &, meansSTPopNorm[[ ;; 65]], meansSTPopNorm[[ ;; 65]]];
```

```
ArrayPlot[PopMatrixSqrt * removingSelfLoops]
```



Link Prediction

From the correlation between two pages and their popularity value, can we say if there is a link or not? We are going to predict the multiplicity of a link, and we are going to assign a error to this prediction.

Distance [u , v] = (1 - corr (u , v))

Adjacency Matrix: 1 / Distances

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

Functions

Prediction Rules

Graph Prediction

Error Measure

Test The Prediction Rules

This prediction rules consider Correlation and Popularity of the pages. If two pages are strongly correlated we put a weighted link between them, proportional to their proximity and the square root of their popularity. Otherwise, if they are not strongly correlated, we consider just their popularity, if their popularity is large, we put a weighted link between them.

```
corrMatrix = Import["/Users/Levantina/Documents/WOLFRAM/PROJECT/Timeseries/
     CorrelationsSTCorrectedTimeSeries.tsv", "TSV"];
```

```
corrMatrixNoSelfCorr =
  corrMatrix * (-1) * (IdentityMatrix[Length[First[corrMatrix]]] - 1);
```

```
corrMatrix65 = corrMatrixNoSelfCorr[[ ;; 65, ;; 65]];
```

```
corrMatrix65 = Import[
   "/Users/Levantina/Documents/WOLFRAM/PROJECT/Timeseries/corrMatrix65.tsv",
   "TSV"];
```

```
meansSTPopNorm = Import[
   "/Users/Levantina/Documents/WOLFRAM/PROJECT/Timeseries/meansSTPopNorm.tsv",
   "TSV"];
```

```
Dimensions[corrMatrix65]
```

```
{65, 65}
```

```
Length[meansSTPopNorm]
```

```
2164
```

```
Export[
  "/Users/Levantina/Documents/WOLFRAM/PROJECT/Timeseries/meanSTPopNorm65.tsv",
  meansSTPopNorm[[ ;; 65]], "TSV"];
```

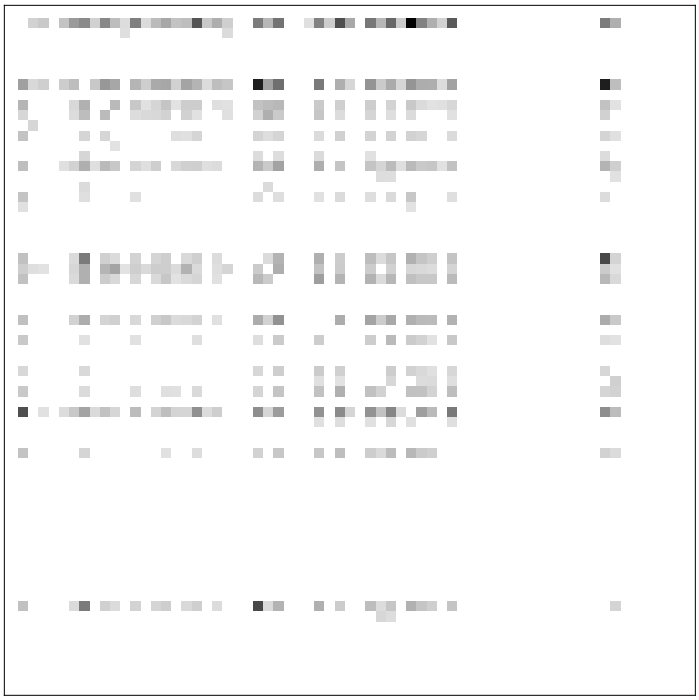
```
/Users/Levantina/Documents/WOLFRAM/PROJECT/Timeseries/meanSTPopNorm65.tsv
```

```
Clear[GraphPrediction1]
```

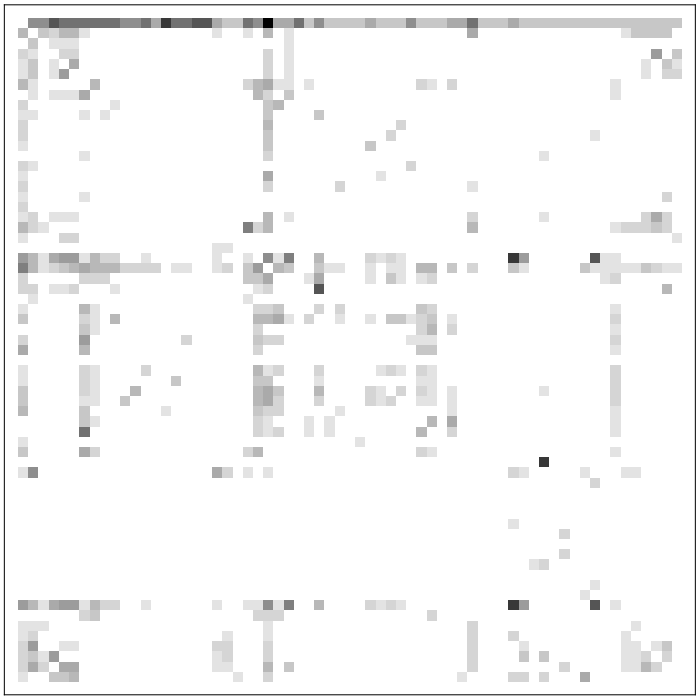
```
predicted = GraphPrediction1[corrMatrix65, Flatten@meansSTPopNorm[[ ;; 65]]];
predictionError[AdjMatr65, predicted]
Mean[Flatten[predicted - AdjMatr65]]
```

```
-0.117821
```

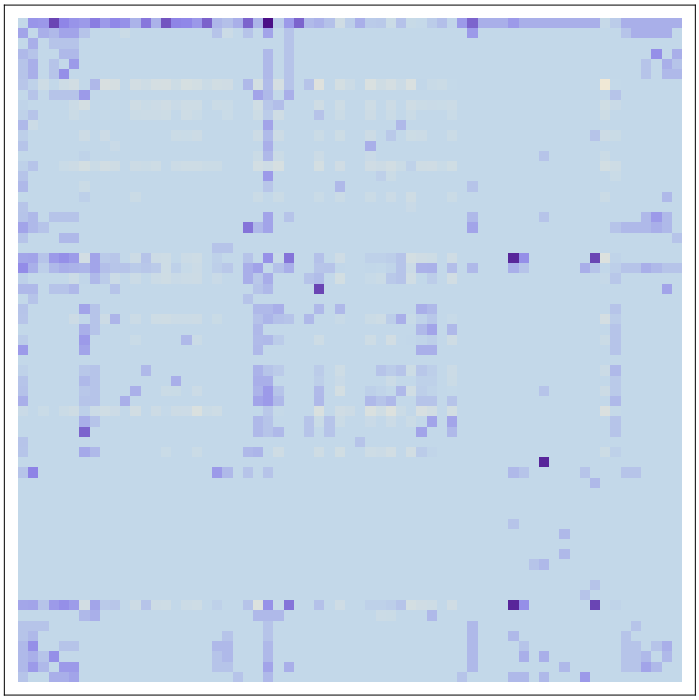
```
ArrayPlot[predicted]
```

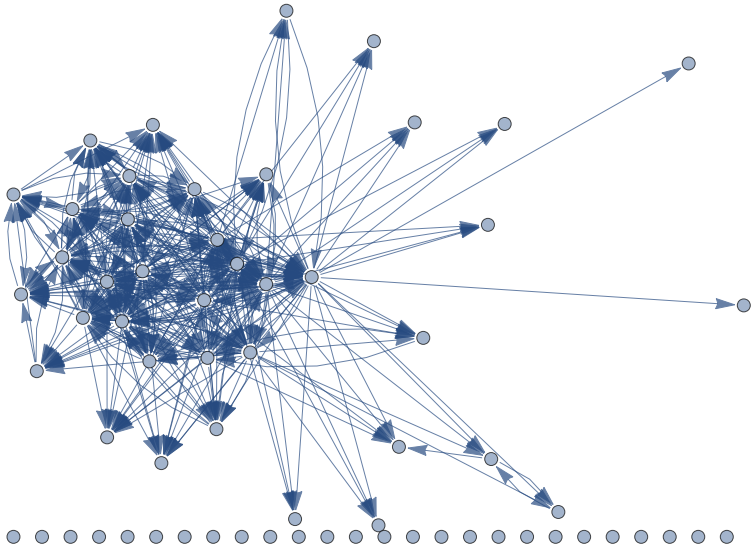


```
ArrayPlot[AdjMatr65]
```



```
ArrayPlot[predicted - AdjMatr65, ColorFunction → "LakeColors"]
```





Live Prediction

Prediction Rule

```
LinkPrediction1[i_,j_,correlationMatrix_, popularities_] := Module[
{predictedLink, a, threshold},
    a = .05;
    threshold = 1.;
    predictedLink = a/(1.-correlationMatrix[[i,j]])*popularities[[i]]^0.5*popularities
    Threshold[predictedLink, threshold]
];
```

Data

```
corrMatrix65 = Import[
    "/Users/Levantina/Documents/WOLFRAM/PROJECT/Timeseries/corrMatrix65.tsv",
    "TSV"];
meansSTPopNorm65 = Import[
    "/Users/Levantina/Documents/WOLFRAM/PROJECT/Timeseries/meanSTPopNorm65.tsv",
    "TSV"];
AdjMatr65 = Import[
    "/Users/Levantina/Documents/WOLFRAM/PROJECT/startrekNetwork/adjMatrix65.tsv"
    , "TSV"];
AdjMatr65 = Partition[Flatten[AdjMatr65], 65];
```

Prediction

```
STVertices[[ ;; 20]]
```

```
{Star_Trek_Into_Darkness, J._J._Abrams, Bryan_Burk,
 Damon_Lindelof, Alex_Kurtzman, Roberto_Orci, Star_Trek,
 Gene_Roddenberry, Chris_Pine, Zachary_Quinto, Zoe_Saldana,
 Karl_Urban, Simon_Pegg, John_Cho, Benedict_Cumberbatch, Anton_Yelchin,
 Bruce_Greenwood, Peter_Weller, Alice_Eve, Michael_Giacchino}
```

```
LinkPrediction1[1, 12, corrMatrix65, meansSTPopNorm65]
```

```
{4.3823}
```

```
AdjMatr65[[1, 12]]
```

```
8
```

Refine the Wikipedia Graph