Further Topics in Social Network Analysis
Friendship paradox in social networks

Dominik Batorski Michał Bojanowski Bartosz Chroł Kamil Filipek

Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw

WARSAW SCHOOL of DATA ANALYSIS

Contents

HUMAN CAPITAL NATIONAL COHESION STRATEGY ICS RUG / UU / RU UNIWERSYTET WARSZAWSKI EUROPEAN UNION EUROPEAN SOCIAL FUND

Project co-financed by the European Union under the European Social Fund

```
##
## Attaching package: 'igraph'

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

1 Introduction

In many, if not in most situations people build their knowledge about society as a whole primarily through their direct experiences from interacting with other individuals. For example:

a) You may asses your professional successfulness by comparing your career to careers of your friends from high school or university. Many people observe that somehow friends tend to be more successful.
b) If you use social media, like Facebook or Twitter, you may have noticed to your discontent, that your Facebook friends, or Twitter followers, seem to have more friends/followers than you do.

First, isn't it paradoxical? Should not the number of friends of friends be, on average, equal to the average number of friends people have?

Second, in all these cases knowledge learned by observing our immediate social environment, i.e. friends, leads to rather pessimistic conclusions regarding our "status" vis a vis others.

In this chapter we will give you some comfort. Social Network Analysis explains why it actually happens to almost all of us. The reason is the so-called *friendship paradox*, first described by Feld (1991). He observed that in friendship networks from US high schools majority of students have less friends than their friends have on average. We also briefly show how this phenomenon, seemingly pessimistic from an individual perspective, can be leveraged to do some good.

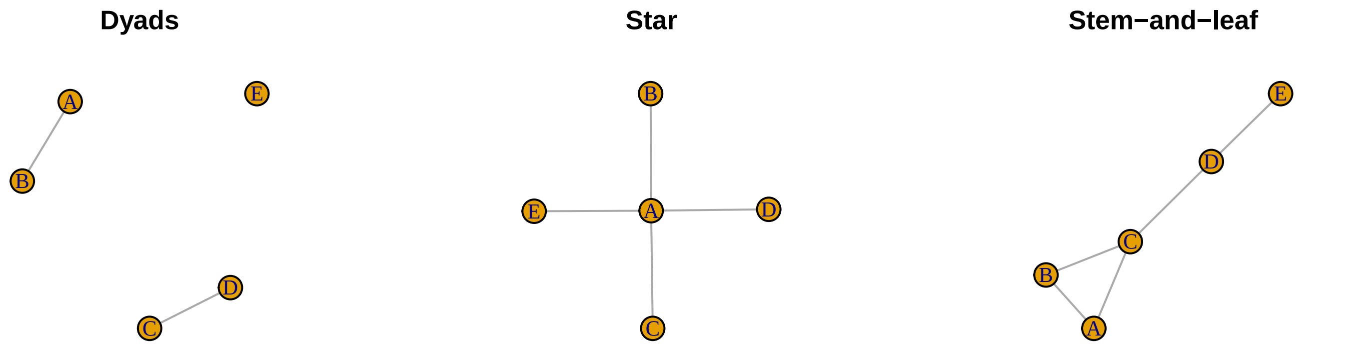
2 Friendship paradox: an anatomy

The most important feature of networks necessary to understand the friendship paradox is degree. In particular, we will need to calculate degrees for individual nodes and compare them to degrees of *friends* of individual nodes.

2.1 Simple examples

Consider the following three simple networks:

```r
library(igraph)
dyads <- graph.formula(A -- B, C -- D, E)
star <- graph.formula(A -- B:C:D:E)
stemleaf <- graph.formula(A:B:C -- A:B:C, C -- D, D -- E)
layout(matrix(1:3, 1, 3))
plot(dyads, main="Dyads")
plot(star, main="Star")
plot(stemleaf, main="Stem-and-leaf")
```



In the `dyads` network nodes A, B, C, and D have all 1 friend each. Node E is an isolate. Average degree is equal to $(4 \times 1 + 0)/5 = 0.8$. If we now look at the number of connections of friends (the number of friends of friends) we will see that:

- A has one friend, B, who has one connection.
- B has one friend, A, who has one connection.
- C has one friend, D, who has one connection.
- D has one friend, C, who has one connection.
- E has no connections.

In consequence, average degree of friends is $(4 \times 1 + 0)/5$ so again 0.8.

In the `star` network A has degree of 4 while B, C, D, and E have degree 1. Average degree is equal to $(4 + 4 \times 1)/5 = 1.6$. If we look at connections of friends we will see that:

- All four friends of A (B, C, D and E) have degree 1.
- B has one friend, A, with degree 4.
- C has one friend, A, with degree 4.
- D has one friend, A, with degree 4.
- E has one friend, A, with degree 4.

Consequently, average degree of friends is $(4 \times 1 + 4 \times 4)/8 = 20/8$ so 2.5, which is greater than average degree of individuals (1.6).

Finally, in the `stemleaf` network:

- C has degree 3
- A, B, and D have degree 2
- E has degree 1

So average degree is equal to $(3 + 2 \times 3 + 1)/5 = 2$.

If we now look at degrees of friends we see:

- A has two friends (B and C) with degrees 2 and 3 respectively.
- B has two friends (A and C) with degrees 2 and 3 respectively.
- C has three friends (A, B, and D) all with degree 2.
- D has two friends (C and E) with degrees 3 and 1 respectively.
- C has one friend (D) with degree 2.

Consequently, average degree of friends is $(2 + 3 + 2 + 3 + 2 + 2 + 2 + 3 + 1 + 2)/10 = 2.2$. Again, greater than average degree of individuals, which was equal to 2.

2.2 Paradox explained

Based on the examples in the previous section we see that average number of friends of individuals (average degree) is never greater, but usually less, than average number of friends of friends. The reason for this seemingly paradoxical fact is that people with many friends, like A in `star` network, or C in `stemleaf` network above, contribute much more to the average degree of friends. Take A in the `star` network:

- A's contribution to average degree is 4.
- A's contribution to the average degree of friends is $4 \times 4 = 16$ because four people mention A as a friend who has 4 connections himself.

In other words, when we calculate average degree of friends, nodes with higher degrees are weighted more than nodes with low degree.

2.3 Friendship paradox: network level and individual level

The *friendship paradox* as described above can be compactly formulated as

> Average number of friends is never greater than the average number of friends of friends.

In this statement we are comparing two quantities

1. Average number of friends.
2. Average number of friends of friends.

These quantities characterize the network as a whole. It can be proved that this statement is always true.

Meanwhile, consider the following statement:

> Average number of friends of friends is greater than the number of friends.

Here we are comparing two quantities:

1. Average number of friends that friends of an individual have.
2. Number of friends an individual has.

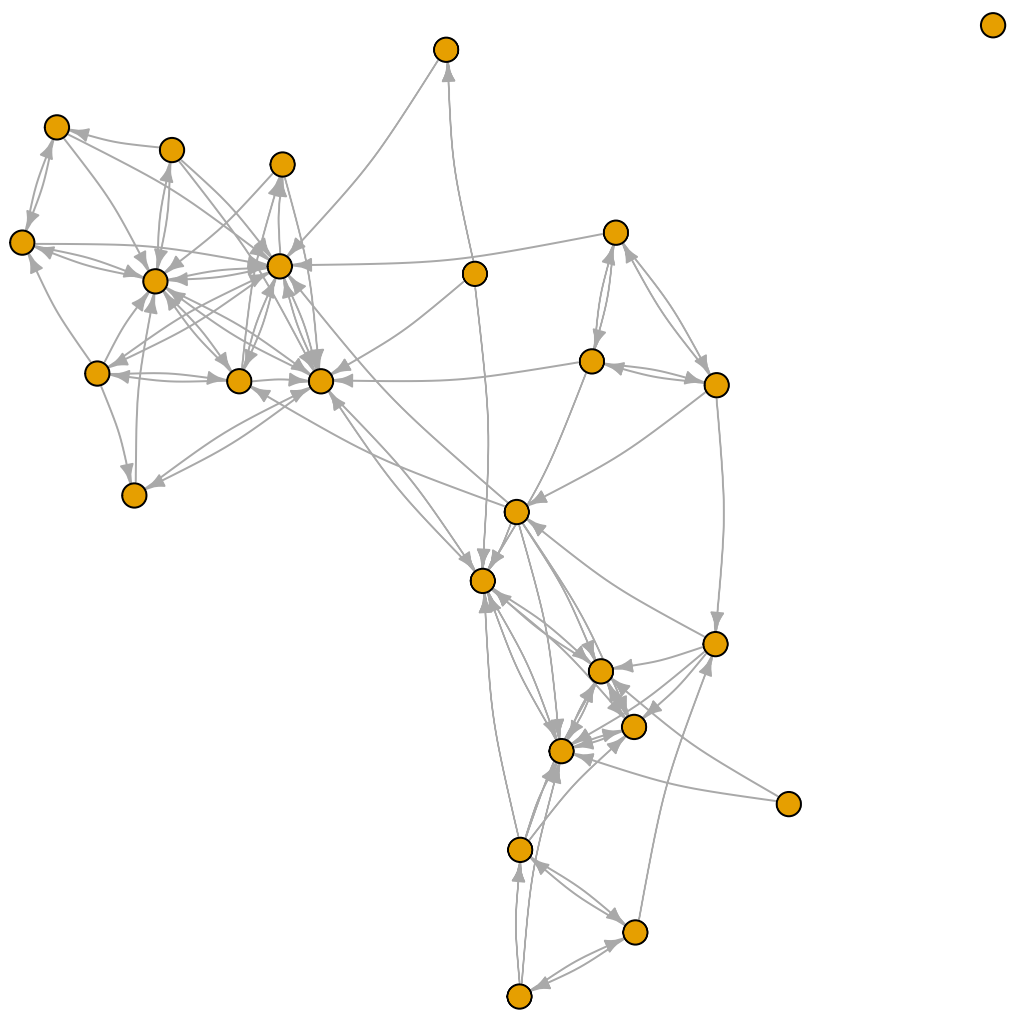
With these statements we are characterizing a particular individual, or node, based on the number of connections he has, and average number of connections his friends have. In other words, these quantities characterize a particular node and its immediate network neighborhood. These are individual-level properties. The statement may hold for some nodes, and not hold for other nodes.

For example, in the `star` network it does not hold for node A as its degree is 4 which is greater than the average degree of friends of A, which is $(4 \times 1)/4 = 1$.

3 Example: classroom network

Let verify the friendship paradox in a real network from a Polish classroom. Children nominated others from their class with whom they like to play with. This is a directed network, so we can analyze in-degree and out-degree separately. Let us focus on in-degree, so "popularity", i.e., how often a kid was nominated by others.

```
library(isnar)
data(IBE121)
playnet <- delete.edges(IBE121, E(IBE121)[question != "play"])
plot(playnet, vertex.size=5, vertex.label=NA, edge.arrow.size=.5, edge.curved=.1)
```



First, we have to calculate in-degrees of all nodes.

```
indegrees <- degree(playnet, mode = "in")
```

Second, we need to identify who are the nominations for each kid. We can obtain it through an *adjacency list*: a list with a component for each kid. Each component is a vector of IDs of nominated kids.

```
adjlist <- get.adjlist(playnet, mode="out")
# entries for first 5 kids
adjlist[1:5]
```

```
## $`1003`
## + 3/26 vertices, named:
## [1] 1018 1051 1072
##
## $`1006`
## + 4/26 vertices, named:
## [1] 1009 1042 1057 1078
```

```
##
## $`1009`
## + 3/26 vertices, named:
## [1] 1006 1057 1078
##
## $`1012`
## + 3/26 vertices, named:
## [1] 1015 1021 1075
##
## $`1015`
## + 3/26 vertices, named:
## [1] 1012 1021 1078
```

From adjacency list we see that, for example, the second kid (number 1006) nominated kids with IDs 3, 14, 19, 26.

Third, we replace node IDs with corresponding in-degrees computed earlier.

```
friends_degrees <- lapply(adjlist, function(id) indegrees[id])
# in-degrees of friends of first 5 kids
friends_degrees[1:5]
```

```
## $`1003`
## 1018 1051 1072
##    9   10    3
##
## $`1006`
## 1009 1042 1057 1078
##    7    9    6    8
##
## $`1009`
## 1006 1057 1078
##    7    6    8
##
## $`1012`
## 1015 1021 1075
##    1    2    2
##
## $`1015`
## 1012 1021 1078
##    2    2    8
```

Finally, we calculate average in-degree among friends (average number of friends of friends) for every node.

```
averages <- sapply(friends_degrees, mean)
averages
```

```
##      1003      1006      1009      1012      1015      1018      1021
##  7.333333  7.500000  7.000000  1.666667  4.000000  5.400000  6.000000
##      1024      1027      1030      1033      1036      1039      1042
##       NaN  7.000000  7.500000 10.000000  5.666667  9.000000  7.000000
##      1045      1048      1051      1054      1057      1060      1063
##  9.000000  6.400000  5.200000  5.600000  7.500000  7.500000  5.000000
##      1066      1069      1072      1075      1078
##  2.000000  4.666667  7.000000  5.750000  6.666667
```

For example, friends of the kid with number 1006 on position 2 have on average 7.5 friends.

Not the kid with number 1024. For him the average number of friends of his friends is `NaN` meaning "Not a Number". He is the isolate in this network: has neither incoming or outgoing connections. Therefore, for him it is impossible to calculate the average.

3.1 Network level

To verify that friendship paradox holds we compare mean in-degree of kids to the mean in-degree of friends of kids:

```
# Average in-degree
mean(indegrees)
```

```
## [1] 3.384615
```

```
# Average in-degree of friends
mean(averages, na.rm=TRUE)
```

```
## [1] 6.294
```

3.2 Individual level

Let us now check how this look on the individual level

```
# For each node,
# is in-degree less than the average in-degree of friends?
# (TRUE or FALSE)
indegrees < averages
```

```
##  1003  1006  1009  1012  1015  1018  1021  1024  1027  1030  1033  1036
##  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE    NA  TRUE  TRUE  TRUE  TRUE
##  1039  1042  1045  1048  1051  1054  1057  1060  1063  1066  1069  1072
##  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
##  1075  1078
##  TRUE FALSE
```

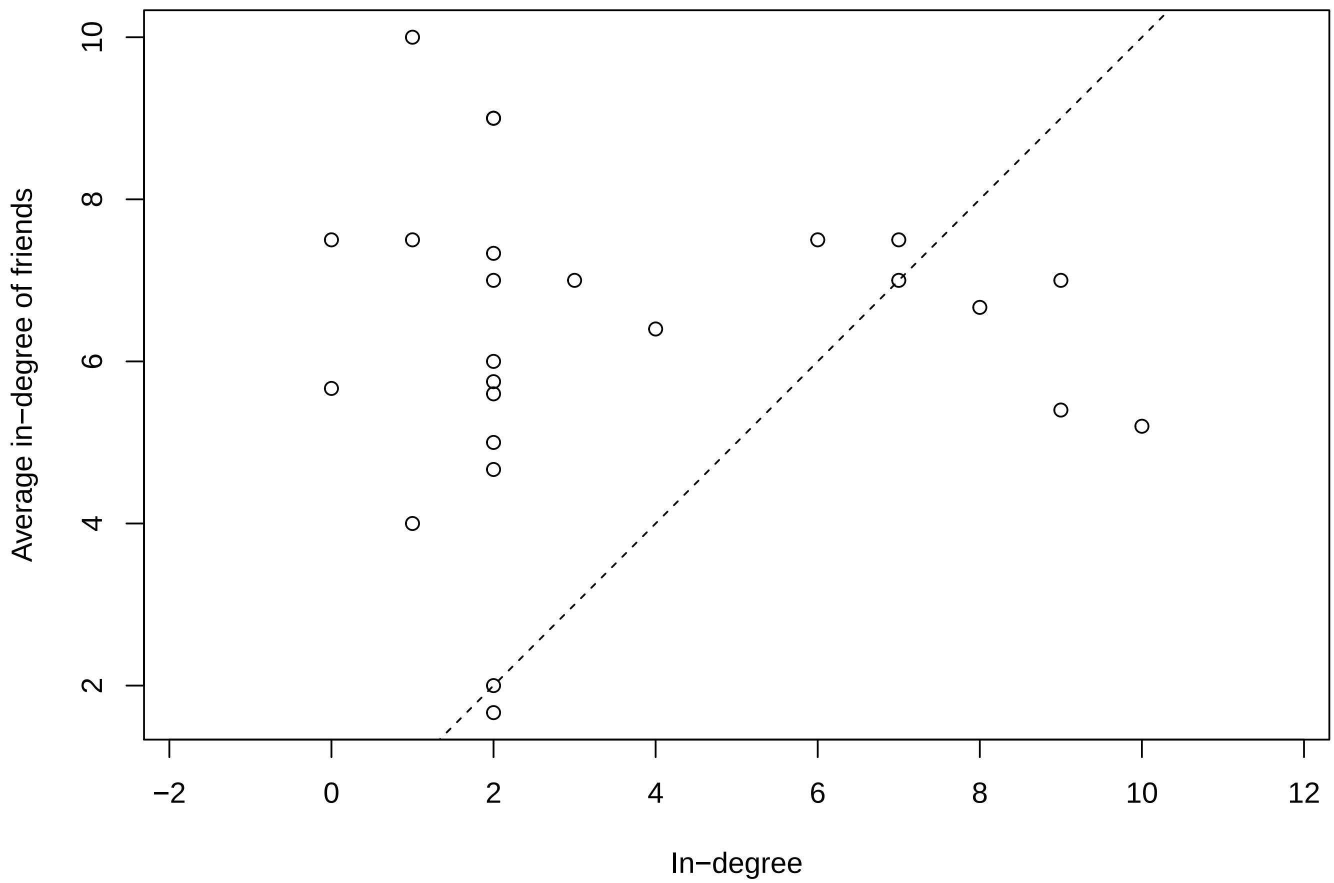
```
# proportion of nodes for which it holds
mean(indegrees < averages, na.rm=TRUE)
```

```
## [1] 0.72
```

Consequently, for 72% of nodes it holds that their friends are on average more popular than they are.

We can also visualize the relationship between popularity (in-degree) and average popularity of friends with the following chart:

```
plot(indegrees, averages, xlab="In-degree",
     ylab="Average in-degree of friends",
     asp=1)
abline(a=0, b=1, lty="dashed")
```



Each point represents a node. The dashed line represent points, for which in-degree is equal to average in-degree of friends. We see that most nodes are above the line, i.e. their friends are on average more popular than they are.

4 Automating computations

For the purpose of demonstration let us create an R function that, when given a network as an argument, will calculate all the quantities we are interested in.

```
friendship_paradox <- function(g, degmode="total", adjmode="total") {
  # compute degrees
  degrees <- degree(g, mode=degmode)
  # identify alters
  adjlist <- get.adjlist(g, mode=adjmode)
  # substitute degrees
  friends_degrees <- lapply(adjlist, function(id) degrees[id])
  # compute avg degrees of alters
  averages <- sapply(friends_degrees, mean)
  # results
  list( average_degree = mean(degrees),  # average degree
       # avg degree of alters
       average_degree_of_neighbors = mean(averages, na.rm=TRUE),
       # prop. of nodes with degree < avg. degree of alters
       p_nodes = mean(degrees < averages, na.rm=TRUE)
       )
}
```

The function requires the argument g with the network. Additionally, it accepts to arguments `degmode` and `adjmode`.

For both acceptable values are "in", "out", or "total". With those two arguments we can define how to calculate degree (degmode) and how to identify network neighbors (adjmode). Be default it calculates total degree and ignores the directionality of the ties. The result is a list with three components:

- average_degree with average degree of the nodes in network g.
- average_degree_of_neighbors with average degree of neighbors.
- p_nodes with proportion of nodes for which their degree is less then average degree of their neighbors.

Let us test it on the classroom network from the previous section:

```
friendship_paradox(playnet, degmode="in", adjmode="out")
```

```
## $average_degree
## [1] 3.384615
##
## $average_degree_of_neighbors
## [1] 6.294
##
## $p_nodes
## [1] 0.72
```

These are identical to our results obtained earlier.

5 Example: Co-authorship network

For the second example let us use a co-authorship network of researchers employed at the University of Warsaw. Nodes are researchers. There is a tie between researchers whenever they have coauthored at least one publication. This network is undirected and rather big – has over 10'000 nodes. Nevertheless, obviously, the friendship paradox still holds: on average coauthors have more coauthors than you do.

```
data(coauthorship)
# number of nodes
vcount(coauthorship)
```

```
## [1] 10114
```

Let us do the computations:

```
friendship_paradox(coauthorship)
```

```
## $average_degree
## [1] 9.716037
##
## $average_degree_of_neighbors
## [1] 16.39974
##
## $p_nodes
## [1] 0.7871268
```

Again, the friendship paradox holds.

Additionally, the percentage of researchers who collaborate less than their coauthors on average is about 87%.

6 Final notes

Do note the difference between the friendship paradox proper, which is defined on the *network level*, and the relation between degree and average degree of neighbors, which is defined on *node level*.

Whether the relationship on the *node level* holds depends on other features of the network, including degree variance, degree mixing (correlations between degrees of neighbors) and so on. See Feld (1991) for more details. Social networks are usually sparse, but clustered, and degree distribution is left-skewed, so the node level relation usually holds.

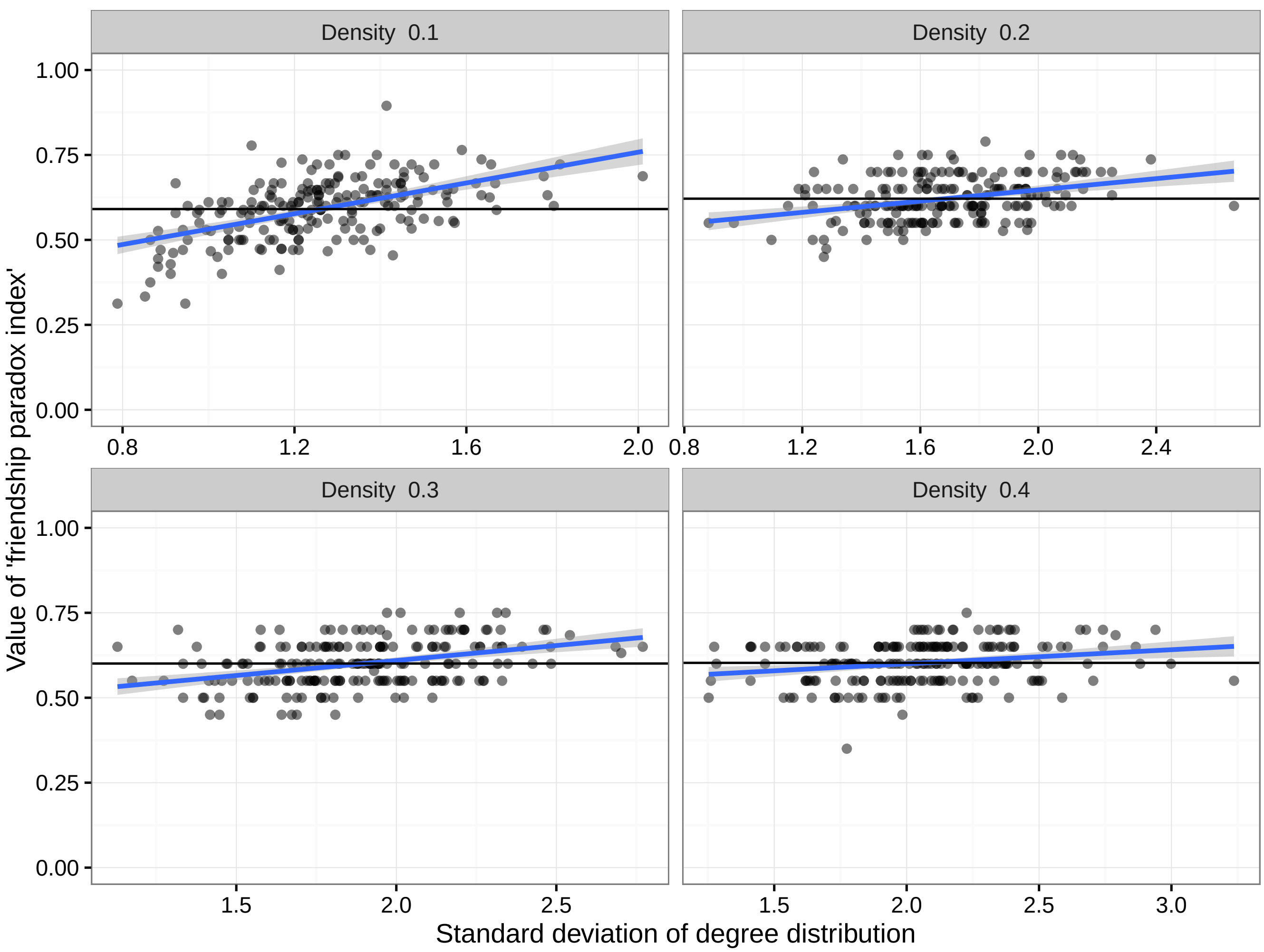
6.1 Friendship paradox and degree variance

Below we illustrate how the proportion of nodes for which it holds that their degree is lower than average degree of their neighbors depends on the standard deviation of degrees in the network. Let us call this proportion "Friendship Paradox Index", or FPI. We do this for four different density levels.

First, we simulate 200 random undirected networks (Erdos-Renyi model) for four levels of density: 0.1, 0.2, 0.3, and 0.4. This gives a total of 800 networks in four density classes. Second, in every such simulated network we calculate:

1. Standard deviation of degrees.
2. Value of FPI.

Results are shown in the figures below. Every point represents a network with a given density, standard deviation of degrees (horizontal axis) and value of FPI (vertical axis).



As you can see, the larger the standard deviation of degrees, the higher the value of FPI. This correlation is stronger for sparser than for lower networks.

6.2 Applications.

For more details and applications see:

- Research of Christakis and Fowler (2010) who propose an efficient vaccination strategy based on Friendship Paradox.
- Hodas, Kooti, and Lerman (2013); Eom and Jo (2014)

7 References

Christakis, Nicholas A, and James H Fowler. 2010. "Social Network Sensors for Early Detection of Contagious Outbreaks." *PloS One* 5 (9). Public Library of Science: e12948.

Eom, Young-Ho, and Hang-Hyun Jo. 2014. "Generalized Friendship Paradox in Complex Networks: The Case of Scientific Collaboration." *Scientific Reports* 4. Nature Publishing Group.

Feld, Scott L. 1991. "Why Your Friends Have More Friends Than You Do." *American Journal of Sociology* 96 (6). The University of Chicago Press: pp. 1464–77.

Hodas, Nathan Oken, Farshad Kooti, and Kristina Lerman. 2013. "Friendship Paradox Redux: Your Friends Are More Interesting Than You." *CoRR* abs/1304.3480. http://arxiv.org/abs/1304.3480.