MA 710 Data Mining

**Furthering Network Analysis**

of

**Ethereum Exchanges’ Transaction Data**

Soubhik Chakraborty

**Spring 2023**

**Abstract** — The Ethereum Network Data Analysis in the group project used only Exponential Random Graph Models (ERGM) for modeling. To distinguish between heavy edges (high value transactions) Vs light edges, we will explore Valued-ERGM (Pavel N. Krivitsky, n.d.) based potential models, which due to hardware resource constraints couldn’t be finished. It is worth looking into it because while vertex attributes driven modeling is very good in explaining the existing dataset, it is not necessarily relevant towards “out of ordinary” anomaly detection and, hence flagging suspicious transactions. A better approach will be to model the overall network shape and interactions using statistical properties of the network itself.

Furthermore, we will see egocentric networks and its potential to do both types of modelling viz. capture per token’s transaction blockchain network properties as well as, overall common nodes across all the tokens cumulative pattern of trades, thus money flow.

1. **INTRODUCTION/MOTIVATION**

**Previous Model Recap:**

A screenshot of a computer code

Description automatically generated with low confidence

One of the key limitations of the above model is for any *vertex: i,j* edges like *j->k* and *k->l* transaction values as weights cannot be used in ordinary ERGM because most methods are node oriented like “nodematch”, “absdiff”.

1. **THE DATA SET**

Towards a fair comparison, exact data is used from the previous effort & advance the existing model such that edges are not lost via sampling, instead use the largest component representative of the “highest” important conglomerate of transactions worth investigating.

1. **MODEL**

Before jumping into model construct, lets briefly look at ergm.control settings.

Note: [Contrastive Divergence](https://deepai.org/machine-learning-glossary-and-terms/contrastive-divergence) maximum iteration is set to 10.

ergm.ctrl = control.ergm(

MCMLE.maxit = 3

**,CD.maxit = 10**

,MCMC.interval = 50

,MCMC.burnin = 40

# ,MCMC.samplesize = 80

# ,MCMC.effectiveSize=20

,MCMC.effectiveSize.maxruns=200

,MCMLE.density.guard=5000

,MCMC.runtime.traceplot=T

**,init.method = "CD"**

,parallel=cores, parallel.type="PSOCK"

,checkpoint="lc1.1.%02d.RData"

# , resume="lc1.3.03.RData"

)

Lets start with a basic edge value model ‘sum’ that uses edge attribute ‘weight’ and references Poisson distribution. As this itself could not be finished in 8hrs time, further modelling is solely based on documentation descriptions of each methods.

enhanced.1 <- ergm(asNetwork(.mgraph) ~ sum + nonzero + mutual("min")

~~+ nodefactor("type")~~

~~+ nodematch("address")~~

**,response="weight", reference=~Poisson, verbose=TRUE**

, control = ergm.ctrl

, san = control.san(SAN.maxit=50))

Because we have seen geometric weighted edgewise shared partners a good fit, lets change this model to use geometric means instead of simple sum. Notice vertex attributes like name2 and type is totally ignored, which is suitable to our purpose. Also, mutual flow is now geometric mean based i.e., if outflow of money is disproportionately different than inflow or vice versa.

enhanced.2 <- ergm(asNetwork(.mgraph) ~ **geomean + nonzero + mutual("geomean")**

, response="weight", reference=~Poisson, verbose=TRUE

, control = ergm.ctrl

, san = control.san(SAN.maxit=50))

Now let’s pick up relevant methods to determine triadic relationships in a more valued manner.

As a token that is being traded amongst multiple stake holders, the valuation of each trade is expected to follow some geometric mean instead of simple sum based arithmetic mean, whereby overall appreciation of a trade is exponential by nature.

enhanced.2 <- ergm(asNetwork(.mgraph) ~ geomean + nonzero + mutual("geomean")

**+ transitiveweights("geomean","sum","geomean")**

**+ cyclicalweights("geomean","sum","geomean")**

, response="weight", reference=~Poisson, verbose=TRUE

, control = lc1.1.ergm.ctrl

, san = control.san(SAN.maxit=50))

Using the equation (13) definition of transitiveweights (Krivitsky, 2012), we can see combine function “sum” is more suitable while value of a “two-path” and final effective pressure is offsetted out between extremes using geometric mean. The modeling is still Poisson based distribution.

Th above models can be tried using Geometric reference as well and observe how mcmc.diagnostics and goodness of fit diagnostics indicate.

1. **MULTINET**

Picking the top 9 tokens most traded in the largest components of 50,000 ethereum transactions, first lets label the tokens with simplistic characters [A-I] thus below:

A picture containing text, screenshot, font, number

Description automatically generatedA screenshot of a computer

Description automatically generated with low confidence

Figure 2

Figure 1

A picture containing text, screenshot, font, number

Description automatically generatedNow having created 9 subgraphs from [A-I], lets look at their isolated token networks’ assortativity w.r.t. exchange type and name2.

A picture containing text, screenshot, number, font

Description automatically generated

Also, lets look at the summary statistics of the same multinetwork.  
**Token F** seems to be quite dense while having just 30 n/w size.

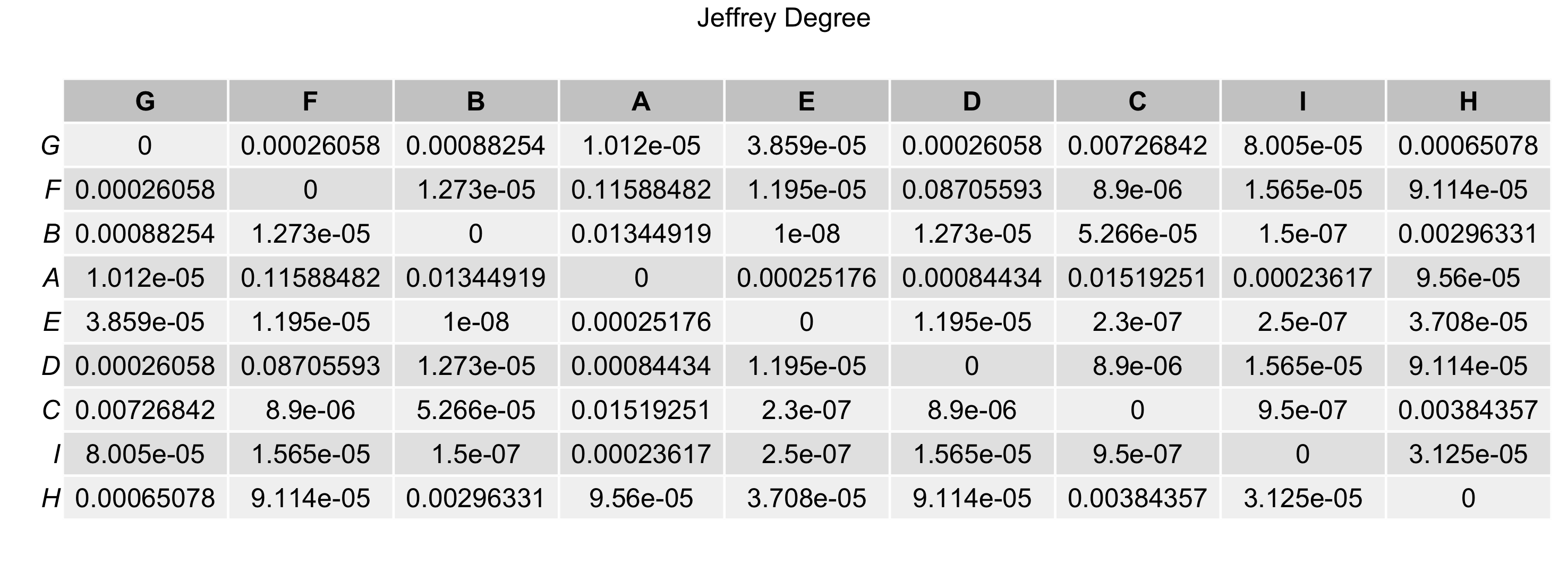
Plotting the multinet shows top row middle column Graph - (F) token is showing **convergence of funds into single exchange DDEX**.

A picture containing line, screenshot, diagram

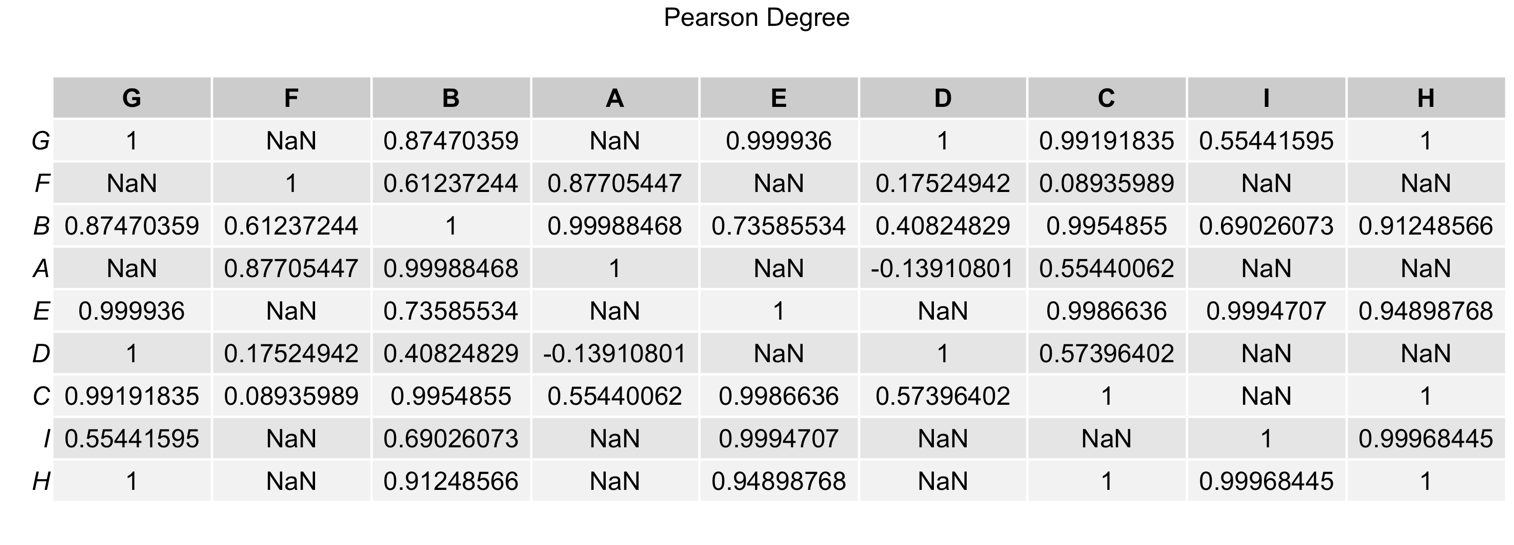
Description automatically generatedA picture containing diagram, pattern

Description automatically generated

Now lets calculate Jeffrey degree and measure Pearson correlation degree among these networks.



We can see token G, E, C,B and A are highly correlated overall transaction history.



A picture containing text, diagram, plan, technical drawing

Description automatically generatedPlotting the degree histogram, we see that D and F have some isolated varying degree, otherwise all other token transaction history is uniform and not showing much disparity in power\_law. Therefore D becomes another **suspicious transaction showing flow of money between Paradex and IDEX**.

A picture containing line, diagram, screenshot

Description automatically generated

A picture containing text, screenshot, number, menu

Description automatically generatedLooking at the exclusive relevance metric which essentially demonstrate how these token networks’ vertices (entities) are highly localised to the exchanges i.e. not present outside its own network layer.

1. **CONCLUSIONS AND FUTURE WORK**

We saw that Valued-ergm has a better promising model and multinetwork statistics already started showing signs of suspicious transactions when we see in parts of the whole and whole divided into parts.

In future, following things is worth trying:

* Try out bootstapped ergms (btergm) and egocentric ermg (egoERGM) on ego.subgraphs
* Lookout for R-GNN and R-GCN implementations in R.

**REFERENCES**

# Bibliography

Krivitsky, P. N. (2012). *Exponential-family random graph models for valued networks*. Retrieved from National Library of Medicine: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3964598/

Pavel N. Krivitsky, C. T. (n.d.). *ERGMs for Valued Networks with Applications to Count Data*. Retrieved from cran-r: https://cran.r-project.org/web/packages/ergm.count/vignettes/valued.html