Gotta Link 'Em All[1]

An Exploration into the Discussion of Pokémon via Social Network Analysis

By: Alex Spancake

Abstract

The world of Pokémon is a complex place, but through the use of social network analysis we attempt to develop an understanding of how individual creatures, referred to as Pokémon, are interconnected. Specifically, we use the frequency in which Pokémon appeared in comments together to reveal potential trends on how users connect these creatures to one another. The 'starter' status along with a Pokémon's type and generation all play some role in how they fit into this simulated social network. Through the use of modularity maximization we define distinct communities and discover characteristics by which they are defined. An ERG model is created but is not predictive. However, it does provide insights on how specific characteristics alter the likelihood of a connection existing between two given Pokémon.

Introduction

On March 29th at 12:58 pm the idea of a Pokémon themed social network project was suggested. I like Pokémon, and I generally prefer not to take tests, so I embarked on a journey that ended up taking way longer than studying for the exam would have. Though as several trainers on Kanto's Route 24 say upon defeat, "I did my best – I have no regrets."[2]

This exploratory analysis aimed at revealing how 502 different Pokémon[3] were connected to one another in the context of general discussion. Specifically, two Pokémon are said to be more strongly connected if they are both regularly mentioned in any given comment. At this point you may wonder, "What kind of characteristics would cause two Pokémon to be perceived as similar?" I'm glad you asked. Below are three examples of Pokémon along with high level characteristics used in this analysis:

Name: Bulbasaur
Type: Grass
Generation: First
Legendary: No
Starter: Yes

Name: Arceus
Type: Normal
Generation: Fourth
Legendary: Yes
Starter: No

Name: Drifloon
Type: Ghost
Generation: Third
Legendary: No
Starter: No

Name – The name of the Pokémon. Some of them are pretty ridiculous.
Type – There are 16 distinct types.
Generation – Every time a new set of Pokémon games are released they include new creatures. Each set of games is referred to as a Generation.
Legendary – Binary; legendary Pokémon are extremely rare, mainly because they cannot reproduce and

[1] A reference to Pokémon's original catch-phrase: "Gotta Catch 'Em All!"
[2] As you play through Pokémon video games you must battle several trainers along the way
[3] There are over 700 Pokémon now, though only those from the first four generations are used in this analysis

hence are the only ones of their kind.
Starter – Binary; a starter Pokémon is one that can be received at the beginning of any game. A Pokémon cannot be both a starter and a legendary.

The final piece of domain knowledge needed before moving forward is that several Pokémon can evolve and thereby form an evolutionary chain. For example:

First Evolution | Second Evolution | Final Evolution

Lastly, it worth noting that since Bulbasaur is considered a starter so are the rest of the Pokémon on its evolutionary chain. These chain lengths range from 1 to 4.

Methodology & Results

1. Data Collection

Nodes

A data set containing basic information of the first 502 Pokémon was used to represent various nodes and their specific characteristics. The 'legendary' and 'starter' variables were manually added.

Edges

Over 300,000 comments were collected from various message boards on Reddit. Specifically, these comments were all made in May of 2015 and are publicly available via Kaggle and can be found here. Once the comments had been downloaded, some text manipulation was done to filter out comments that did not include any of the 502 valid Pokémon.

At this point the remaining comments were broken apart into several strings of individual words. If a given word matched the name of a Pokémon then it remained, else it was removed. Luckily people are more meticulous than you may imagine when it comes to correctly spelling Pokémon names. From a random sample of 200 comments I only found a single comment with an incorrectly spelled Pokémon name. Unfortunately I accidently filtered out Mr. Mime at this point since his name is two separate words, sorry man.

Next, undirected edges were created. For example, if Bulbasaur was mentioned in the same comment as Arceus on four separate occasions, then an edge with weight four was created. This created a lot of edges. The vast majority of edges had a strength value of 1. In order to filter down to connections that were meaningful, only edges that had a strength of 3 or greater were included in the network. This also reduced the number of nodes that were included in the analysis since several Pokémon were not mentioned in this subset of comments.

With both a node and edge list in hand, the data was prepared for some network analysis.

2. Network Descriptives

Attempting to understand some of the fundamental features of this network seemed like a reasonable place to start. For example, this network had an **average degree** of **3.7**. In other words, your average

node had 3 or 4 associated edges. This does not take into account the **average strength** of edges, which average about **7** and formed the following distribution:

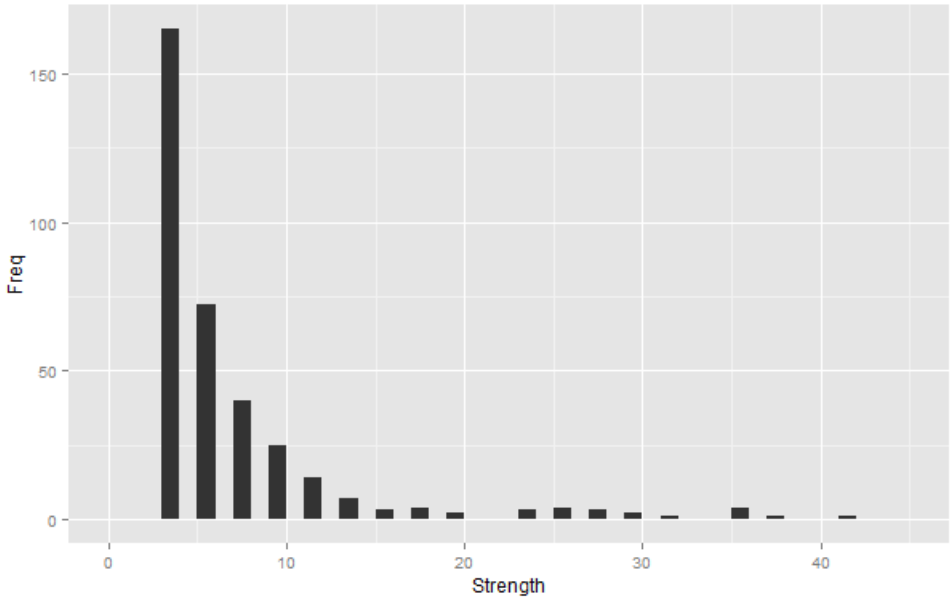


Figure 1. Distribution of Comment Strengths

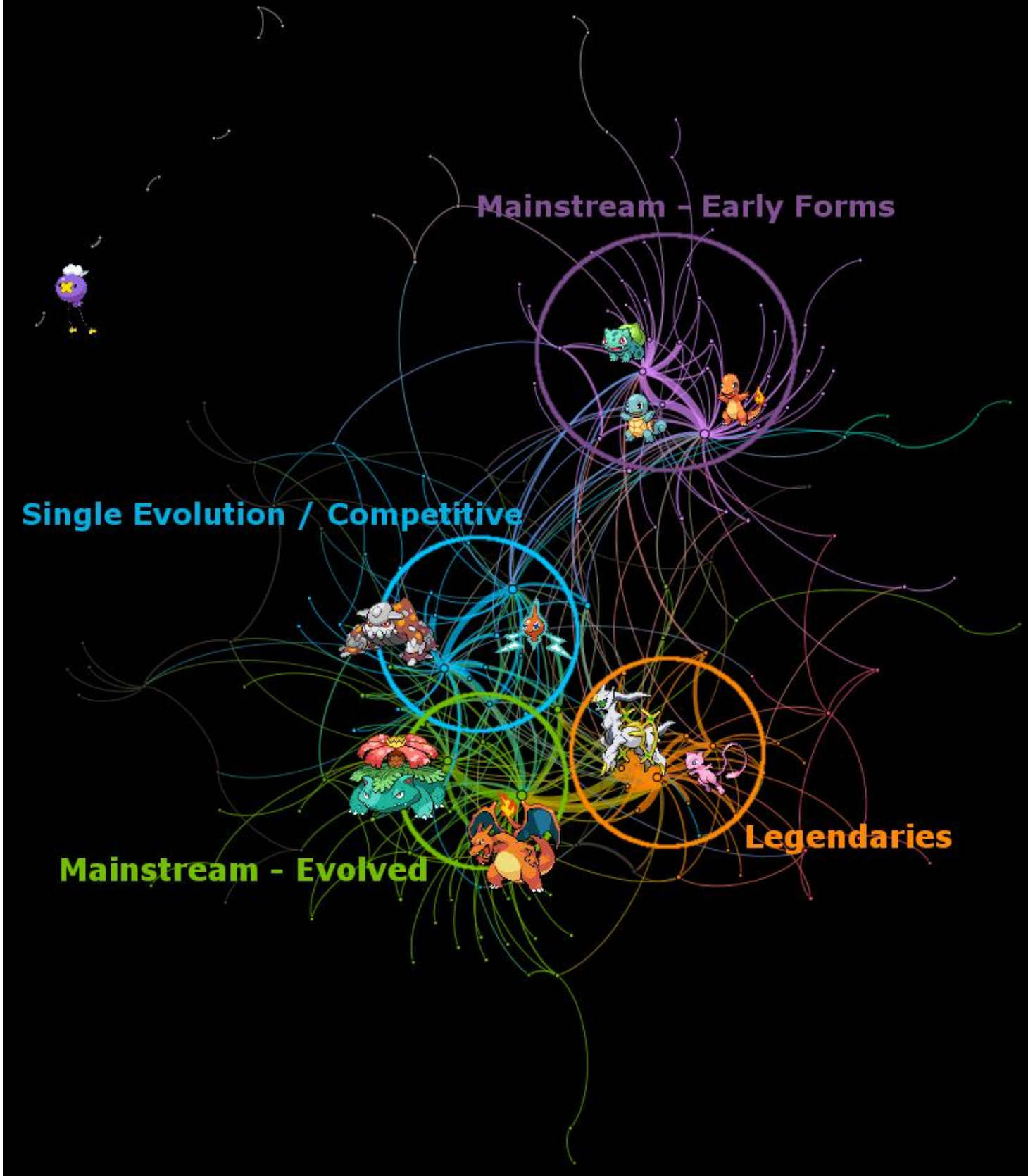
Take note of the right-skewed distribution. Although we have limited edges to only those which had at least a strength of 3, it appears that a small subset edges represent the defining features of this network.

Finally we move on the measuring the **density** of this network which represents how interconnected the various nodes are. Specifically, this network contains only **1.9%** of all possible connections that could feasibly occur. If all of the original 502 nodes were included, this value would be even lower.

3. Modularity

I knew that at some point throughout this project I wanted to make a pretty picture, so I started taking a closer look at various clustering algorithms developed for graphs. Modularity maximization is currently the most popular algorithm, and more importantly it was the easiest to implement. This network received a **modularity** of **.524.** This value could have been as low as -.5 (this network wasn't much different than a network of randomly placed edges) or as high as 1 (this network definitely has unique communities.) In this context, it appeared that the network had the potential to be broken up into communities for further investigation. Modularity maximization yielded a total of 14 communities. In order to visualize them, each community was assigned a distinct color. Additionally, a force-directed graph drawing algorithm (Force Atlas) was used to make the following visualization more aesthetically pleasing.

Social Nework Analysis - Pokemon Edition!



The thickness of each edge corresponds to its relative strength. The size of each node is based on its eigenvalue centrality. This metric was chosen over the node's degree because it accounts for the relative importance of a given node's contacts, in addition to itself. As noted on the graph, four major communities became apparent.

Mainstream – Early Forms: These Pokémon are a combination of well-known Pokémon who are often the first evolution of a larger chain. The three nodes with the highest eigenvalue centrality were the three starter Pokémon in the original series and are some of the most recognizable to children/young adults across the country.

Mainstream – Evolved: Similar to the previous community, these Pokémon are some of the most recognizable. The biggest difference is that they are almost all fully evolved. Like the first group, the nodes with the largest eigenvalue centrality are also starter Pokémon from the first generation.

Legendaries: As mentioned earlier, these Pokémon are extremely rare and hard to obtain. Often times their backgrounds are interconnected. This community was almost exclusively legendary Pokémon. The node with the highest eigenvalue centrality was Arceus, who created and controls the Pokémon universe.

Single Evolution / Competitive: This group is the least defined of the four, much of the interpretation is speculative. However, a trend that I noticed was that these Pokémon are hard to obtain and are often part of an evolution chain with length 1. Furthermore, they are strong on the battlefield and are commonly discussed amongst trainers that battle competitively. The node with the highest eigenvalue centrality is Heatran.[4]

A final note – a weakness of modularity maximization is that each node can only be a part of a single community. As evidenced by the somewhat vague descriptions of the most prevalent communities, many Pokémon could be part of more than one community. However, a different clustering technique would have to be used in order to reveal these trends.

4. ERG Modeling

With an idea of how nodes were connected, I decided to take a stab at ERG modeling to determine if the features collected about each Pokémon could be useful in predicting the probability of a link existing between any two nodes. Through much trial and error, the following variables proved significant in predicting the probability of an edge existing with an alpha of 5%. A mixture of standard terms and homophily terms were used to create various models. Models were compared via AIC and the number of triangles. The following variables created the 'best' model.

Variable	Type	Estimated Coefficient	p-value
Edge	Intercept	-6.96	< .001
Dark Type	Factor	2.28	.003
Ice Type	Factor	1.89	.007
Fourth Gen	Factor	-1.24	< .001
Second Gen	Factor	-1.56	< .001
Third Gen	Factor	-1.31	< .001
Type	Homophily	-1.28	.018
Starter	Factor	3.12	< .001

Table 1. ERGM Variables – Final Model

[4] Heatran is actually a dual type Pokémon. Specifically, he is both Fire and Steel. This unique typing allows him to strongly resist most attacks, hence making him a strong competitive choice.

These coefficients yield some important results. First, it appears that some Pokémon types have, on average, more connections than others. Both dark and ice types contain some pretty 'cool' Pokémon, though that may just be a matter of personal opinion. It does appear that starter Pokémon are likely to form more connections. Furthermore, Pokémon from the first generation rank as the most likely to form additional edges compared to their second, third, and fourth generation counterparts. Lastly, and most surprisingly, the homophily term for Pokémon sharing the same type resulted in a negative coefficient. My gut is that sharing the same type does not make a significant difference on the probability of developing an edge, though this would need to be formally tested.

While this ERGM analysis provided some potential insight on how various features affect the probability of developing additional edges, it is exploratory at best. Unfortunately I do not recommend using it as a predictive model either. Simulated networks using this model do not come close to resembling the actual network with regards to the number of triangles produced.

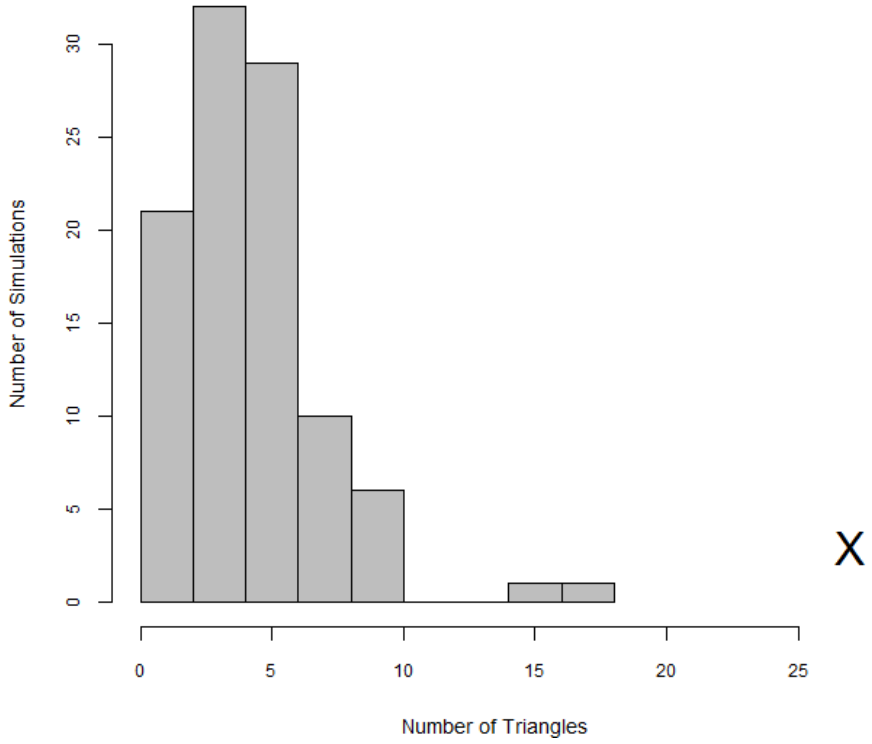
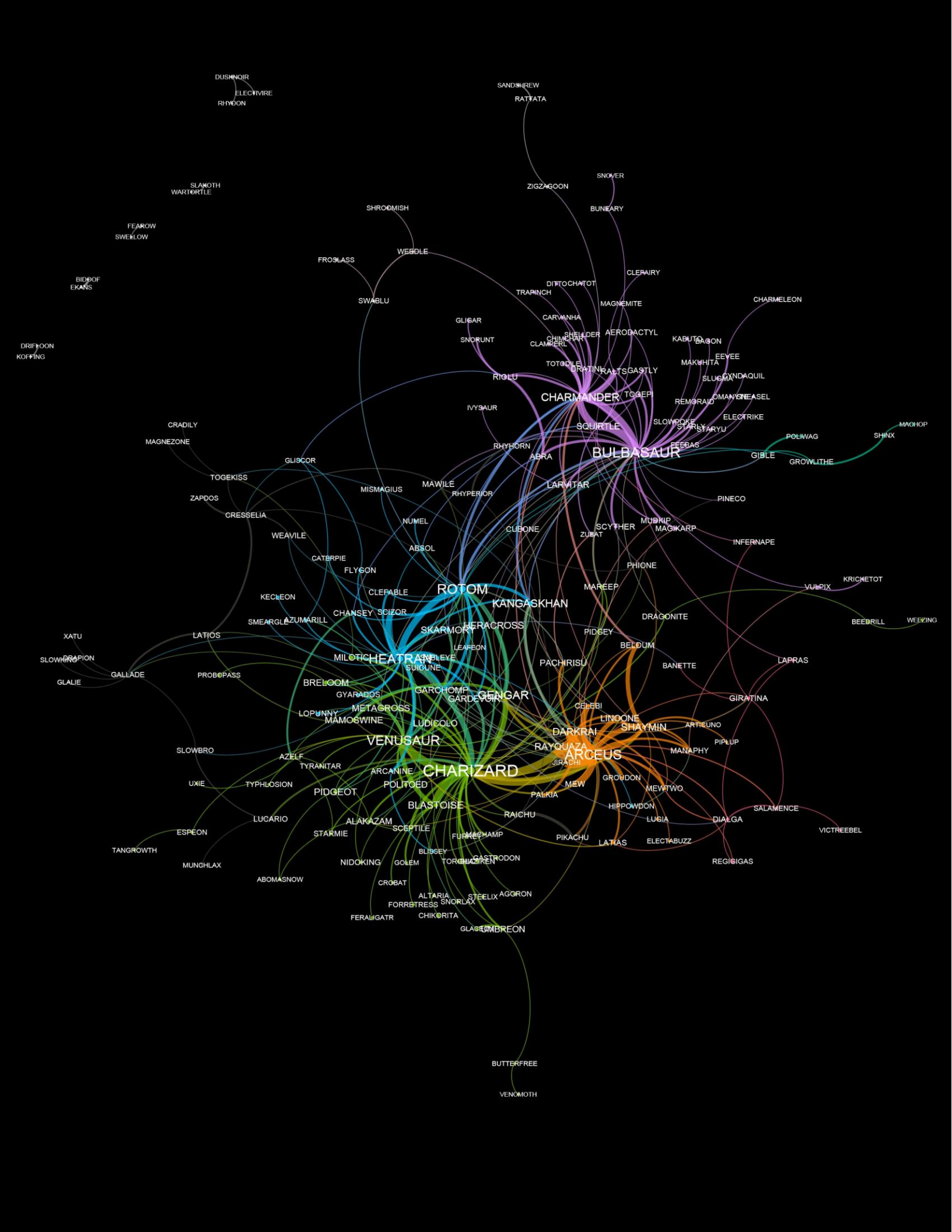


Figure 2. Number of Triangles – Simulated vs Actual

Conclusion

The world of Pokémon contains all sorts of strange complexities would take a long time to understand. That being said, we use the frequency in which Pokémon appeared in comments together to reveal definitive trends on how users connect individual creatures to one another. The starter status, typing, and generation of a Pokémon all play some role in how they behave in our simulated social network. The ERG model created is not predictive but does provide some room for further investigation. Specifically, the following appear to be potential avenues for further research:

- Utilizing hypothesis testing to formally test the significance of characteristics such as generation
- Are additional variables such as stage within an evolutionary explain network connections?


DUSKNOIR
ELECTIVIRE
RHYDON
SANDSHREW
RATTATA
SNOVER
ZIGZAGOON
SLAKOTH
WARTORTLE
BUNEARY
SHROOMISH
FEAROW
SWELLOW
WEEDLE
FROSLASS
CLEFAIRY
BIDOOF
EKANS
DITTOCHATOT
TRAPINCH
SWABLU
MAGNEMITE
CHARMELEON
GLIGAR
CARVANHA
SHELLDER
AERODACTYL
SNORUNT
CHIMCHAR
CLAMPERL
KABUTO
BAGON
DRIFLOON
KOFFING
EEVEE
MAKUHITA
TOTODILE
DRATINI
RALTS
GASTLY
SLUGMA
CYNDAQUIL
RIOLU
CHARMANDER
TOGEPI
REMORAID
OMANYTE
SNEASEL
IVYSAUR
ELECTRIKE
CRADILY
SQUIRTLE
SLOWPOKE
STARLY
STARYU
MAGIKOP
MAGNEZONE
POLIWAG
SHINX
RHYHORN
FEEBAS
BULBASAUR
ABRA
GIBLE
GLISCOR
GROWLITHE
TOGEKISS
MISMAGIUS
MAWILE
LARVITAR
RHYPERIOR
ZAPDOS
PINECO
CRESSELIA
NUMEL
MUDKIP
SCYTHER
CUBONE
MAGIKARP
ZUBAT
WEAVILE
INFERNAPE
ABSOL
CATERPIE
PHIONE
FLYGON
MAREEP
KRICKETOT
ROTOM
VULPIX
CLEFABLE
KECLEON
KANGASKHAN
CHANSEY
SCIZOR
DRAGONITE
SMEARGLE
AZUMARILL
BEEDRILL
WEEZING
SKARMORY
HERACROSS
XATU
LATIOS
PIDGEY
LEAFEON
BELDUM
SLOWKING
DRAPION
MILOTIC
HEATRAN
ASHLEYE
SUICUNE
PACHIRISU
LAPRAS
GALLADE
PROBOPASS
BANETTE
GLALIE
BRELOOM
GYARADOS
GARCHOMP
GENGAR
GARDEVOIR
GIRATINA
METAGROSS
CELEBI
LOPUNNY
MAMOSWINE
LINOONE
SHAYMIN
ARTICUNO
LUDICOLO
DARKRAI
VENUSAUR
PIPLUP
SLOWBRO
RAYQUAZA
MANAPHY
AZELF
ARCEUS
TYRANITAR
JIRACHI
ARCANINE
CHARIZARD
GROUDON
UXIE
TYPHLOSION
POLITOED
MEW
MEWTWO
PIDGEOT
PALKIA
BLASTOISE
HIPPOWDON
SALAMENCE
RAICHU
LUCARIO
LUCIA
DIALGA
ALAKAZAM
SCEPTILE
ESPEON
STARMIE
PIKACHU
VICTREEBEL
TANGROWTH
LATIAS
ELECTABUZZ
BLISSEY
NIDOKING
GOLEM
CASTRODON
REGIGIGAS
MUNCHLAX
CROBAT
ABOMASNOW
ALTARIA
STEELIX
AGGRON
FORRETRESS
SNORLAX
CHIKORITA
FERALIGATR
GLACEON
UMBREON
BUTTERFREE
VENOMOTH