Analysis of anime voice actor's social network and popularity

Florencia Zanollo - Hideaki Takeda

University of Buenos Aires (Argentina) - National Institute of Informatics (Japan)

Introduction

Our research went through multiple stages. Starting with Linked Open Data, moving to Social Network Analysis and ending with applied machine learning.

Thus the order of this presentation will be chronological.

Nonetheless our main focus is: *Social network of seiyuu* (or anime's voice actors)

Why seiyuu?

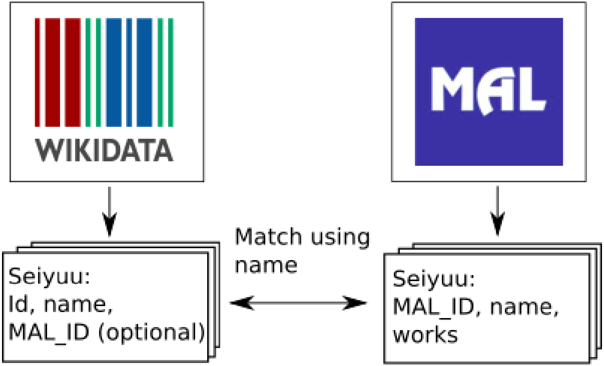
- Because I **really** like anime and manga.
- There are many researches whose focus is actor's social network but usually is hollywood actors and not only voice actors. Seiyuu and anime industry is really unusual and so could present a different structure.
- There're several database with information about anime and seiyuu but either it's incomplete or doesn't have a good structure nor format.

Wikidata and MyAnimeList

We wanted to use *Wikidata* as our source but it's too incomplete; it doesn't have information about works of seiyuu.

Instead we used *MyAnimeList* (MAL) through an API called *Jikan*; retrieving data in JSON and then changing its format to RDF.

We used Wikidata to get the list of seiyuu and MyAnimeList to get list of works for each seiyuu and information about anime.



- Total of 6472 seiyuu on Wikidata
- Only 59 had MAL_IDs
- 3033 MAL_IDs retrieved
- At the end having 3092 seiyuu with MAL_ID
- 2956 of which had at least one work

Data retrieved

All in all we were able to retrieve the following information for *2956 seiyuu and 7614 anime*.

- For Seiyuu:
    - Name
    - Debut (this was obtained from oldest work's aired date)
    - Gender
    - Popularity (member_favorites information of MAL)
    - Works (anime roles with anime information plus wheter is a main role or not)
- For Works (Anime):
    - Year that began airing
    - Favorites
    - Score (from 0 to 10, MAL user based)
    - Popularity (ranking over all MAL animes)
    - Members (how many MAL users have it on their list)
    - Genres

Social Network

This social network is of a particular kind called *two-mode networks* which consists of a set of actors (seiyuu) and events (anime).

Details to consider:

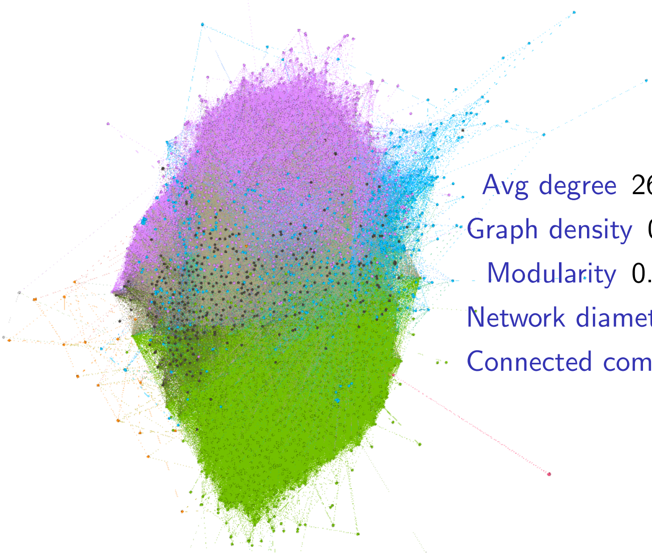
- We can choose between anime and seiyuu as nodes.
- Nodes are time dependant (since they have debut year).
- Edge or relationship definition:
  - How many works in common?
  - Which time frame?

We compared two graph definitions using *seiyuu as nodes* and, as edge definition:

- at least 1 work in common
- at least 10 works in common

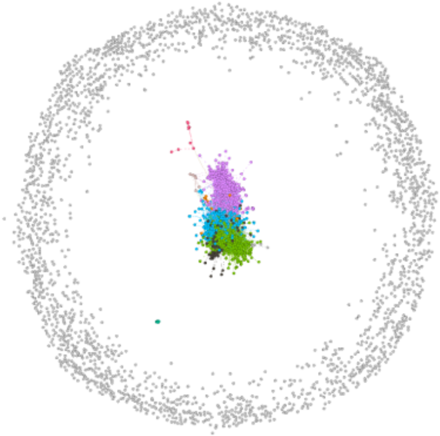
Both of them during the time frame between the first debut registered (1960) and the year of observation.

One work in common



Avg degree	267
Graph density	0.09
Modularity	0.2
Network diameter	6
Connected components	18

Ten works in common



Avg degree 9
Graph density 0.003
Modularity 0.29
Network diameter 7
Connected components 2261

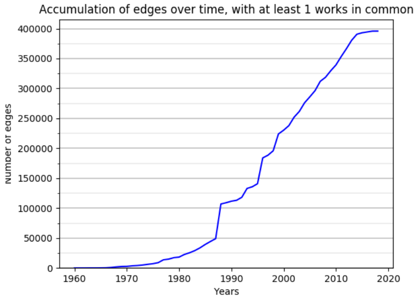
Features

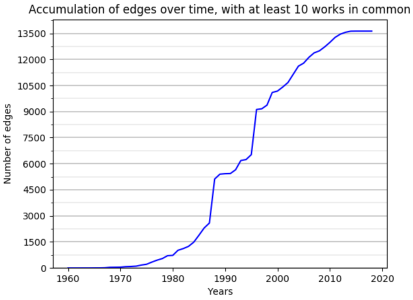
Strongly connected With merely ∼3000 nodes it has ∼400000 edges when only one work in common is required and ∼14000 edges when asking for 10 or more.

Big cluster Thightly interconnected and big cluster surrounded by poorly or not connected nodes. (99% of the nodes of one work in common graph and 23% of 10 works in common).

Communities We can see at least four clear communities in each graph.

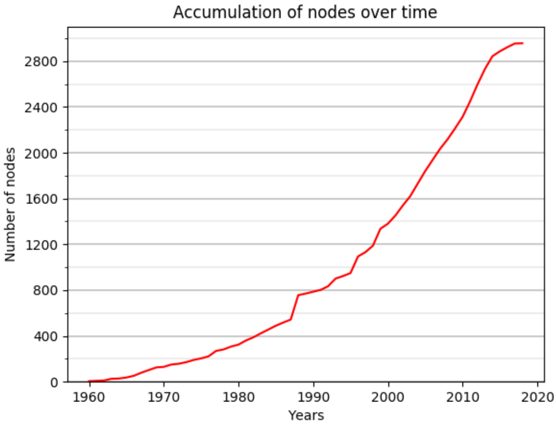
Edge growth





- Edge growth follows the same distribution with 1 and 10 works in common

Node growth



- More than half of the nodes are from last 18 years (2000 to 2018)

Top 5 degree

Name	Degree
Takehito Koyasu	1545
Akira Ishida	1488
Mamiko Noto	1422
Nobuo Tobita	1417
Daisuke Namikawa	1390

Table: One work in common

Name	Degree
Takehito Koyasu	311
Akira Ishida	273
Mamiko Noto	258
Daisuke Namikawa	232
Katsuyuki Konishi	229

Table: Ten works in common

Top 5 betweenness centrality

Name	BtwC
Takehito Koyasu	49982.52
Akira Ishida	40221.50
Daisuke Namikawa	30448.43
Nobuo Tobita	29363.25
Mamiko Noto	29168.18

Table: One work in common

Name	BtwC
Takehito Koyasu	18489.44
Mamiko Noto	10988.96
Daisuke Namikawa	9570.48
Akira Ishida	8299.19
Rie Kugimiya	7560.16

Table: Ten works in common

Definition

Both networks have fairly similar top 5s so it points to them having similar structure and connections amoung their nodes, aside from actual values.

From now on our graph definition will be:

Node Seiyuu

Edge At least 10 works in common from 1960 to year of observation.

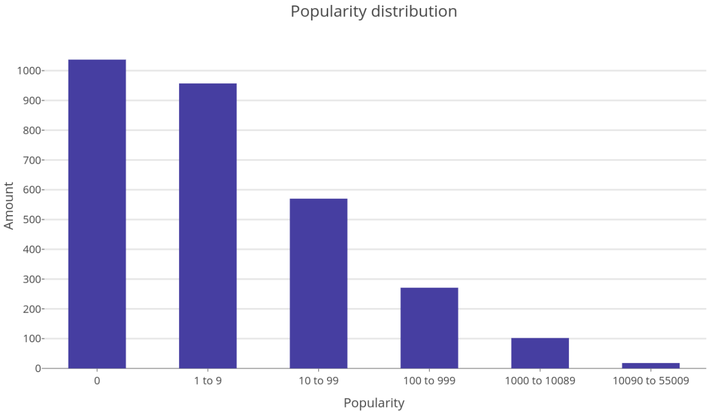
Because requiring more jobs in common means less amount of edges, this leaves a more understandable graph and we verified it does without changing its structure so much.

Popularity: Definition

Since we are using MAL database and it has a social component, seems logic to use *member_favorites* as a representation of *popularity*. We can also get popularity and score of anime from opinions of the same set of users.

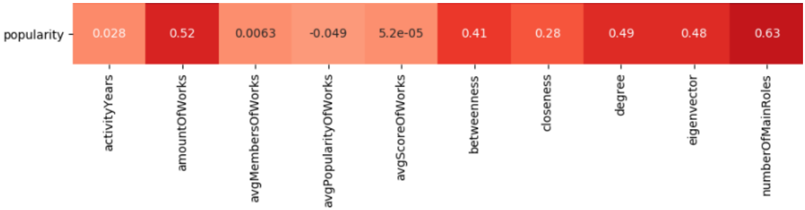
Name	Popularity	Some popular roles of them
Kana Hanazawa	56637	*Angel Beats!*: Tachibana, Kanade
Hiroshi Kamiya	49685	*Shingeki no Kyojin*: Levi,
Mamoru Miyano	43942	*Death Note*: Yagami, Light
Rie Kugimiya	31668	*Fullmetal Alchemist*: Elric, Alphonse
Jun Fukuyama	26811	*Ao no Exorcist*: Okumura, Yukio
Miyuki Sawashiro	26501	*Durarara!!*: Sturluson, Celty
Tomokazu Sugita	24449	*Gintama*: Sakata, Gintoki
Daisuke Ono	24080	*Durarara!!*: Heiwajima, Shizuo
Saori Hayami	18322	*Owari no Seraph*: Hiiragi, Shinoa
Aya Hirano	18094	*Fairy Tail*: Heartfilia, Lucy

Popularity: Analysis



- Total: 2956
- Mean: 289.55
- Median: 2.0
- Min: 0; Max: 55018
- 1037 values equal to zero
- Only 120 values bigger than 1000

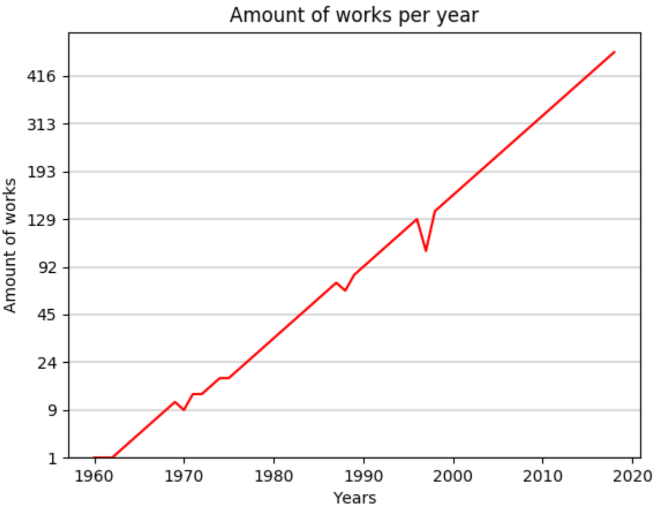
Popularity: Pearson correlation



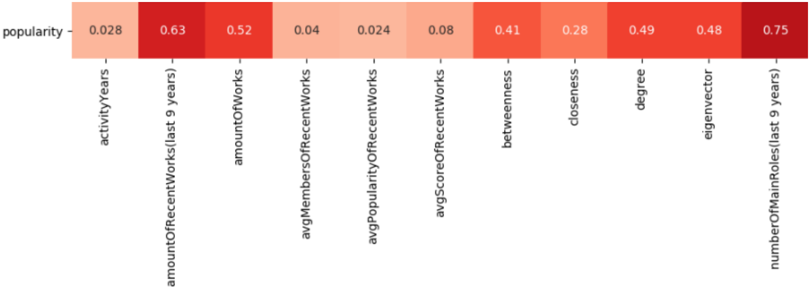
- Big correlation between popularity and amount of works.
- Number of main roles and amount of works have a strong correlation with each other (0.9) but they have different influence over popularity, this means they provide distinct information.

Since our dataset is biased in favor of more modern anime we thought of correlate with *recent works* only.

But, how recent? Last 5, 10 or 20 years?



So our definition of recent works will be last 9 years.



- Amount of recent works has more correlation with popularity than not recent.
- It also happens with number of main roles.

Next graphs were made in order to discover *why 9 years* prior had more correlation.

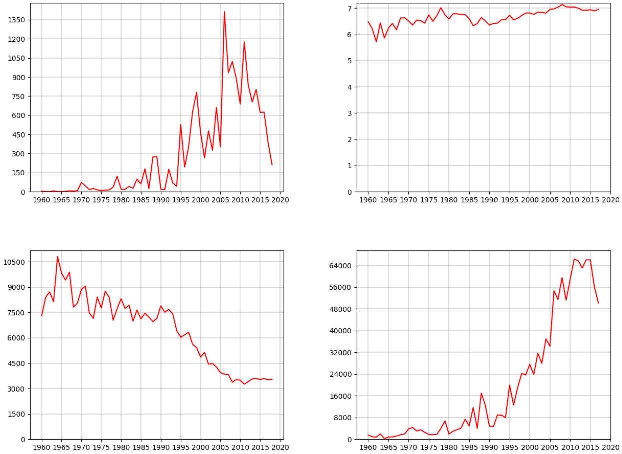


Figure: Averages of favorites, score, popularity and amount of members.

Popularity: Prediction

The node attributes were divided into categories, leaving four distinct types:

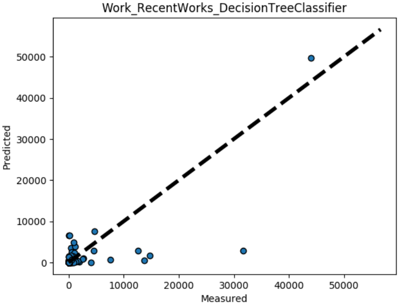
- Personal data:
  - Debut
  - Gender
  - Activity years (2018-debut)
- Works data:
  - Amount
  - Top 5 genre
  - Favorites
  - Score
  - Popularity
  - Members
  - Number of main roles
- Recent works data:
  - Same as works but for only last 9 years
- Graph data:
  - Degree
  - Betweenness centrality
  - Closeness

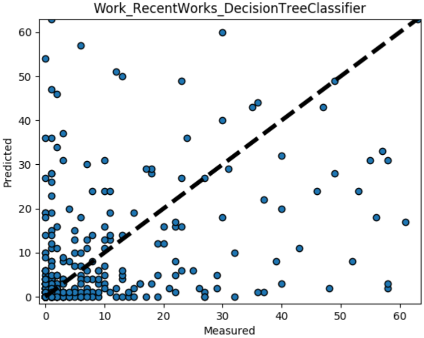
Fitting and prediction experiments were run for each category, each combination of 2, 3 and all of them together; using Scikit-learn, 80% of seiyuu as train data and the rest as test. This was done for all following models:

- DecisionTreeRegressor
- DecisionTreeClassifier
- LinearRegression
- KNeighborsClassifier
- LinearDiscriminantAnalysis
- GaussianNB
- SVM

Because popularity variance is really high we got good results in metrics but particular predictions were aloof.

For example:
DecisionTreeClassifier gets a r2_score of 0.57 when using model data from *works and recent works* categories but the scatter plot doesn't show such good results.



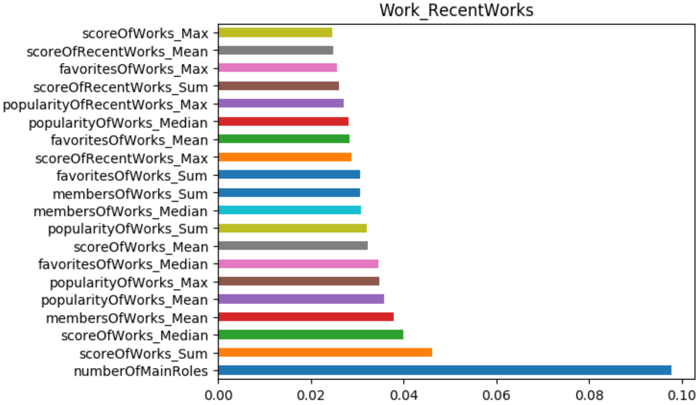


(a) Zoomed in

Popularity: Feature importance

One of our goals about predicting popularity was to use the classifier algorithms to understand which are the key features that differentiates between a "not so much" and a popular seiyuu.

For our previous example we got:



Conclusion

Since particular predictions were usually wrong (mostly for small values) this couldn't be used in a practical way to predict popularity of new seiyuu nor to recover lost values.
But we gain knowledge of

- state of the art of seiyuu and anime data; which is very incomplete or withouth good format.
- structure and characteristics about seiyuu's social network.
- better understanding about what is it that makes a seiyuu popular in terms of MAL users.

Future works and Thanks!

Future works:

- Compare social network growth with particular graphs to see in which category it fits.
- Find a better metric to compare popularity prediction results.
- Improve models and data for better popularity predictions.
- Organize seiyuu and anime data using Wikidata ontology and make it open.

Thank you for your attention, feel free to ask questions!