Fundamental structures: Many-to-many asynchronous

Benjamin Skjold*[1]

[1] *Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Building 321, Technical University of Denmark, 2800 Lyngby, Denmark*

June 28, 2021

1 Introduction

In the recent years an increasing amount of the communication between people has been moving online to social platforms, such as Facebook, Twitter, Reddit and many more. This has been heavily studied in the light the language used. However, we are interested in another aspect of these communication events, namely the social networks that are emerging from the users interactions. In Lehmann (2019) we introduce a framework for modelling temporal communication networks, where we divide networks into meaningful classes based on its generating process. We classify the networks emerging from online communication as a many-to-many asynchronous network. This is by no means the first study that investigates these networks (see eg. Gómez et al. (2008); Thukral et al. (2018)). However, these have simply chosen an interesting measure for that given network and investigated that. I highlight that this is an abuse of the word ”simply”, as the work done is by no means simple. However, we want to take a step back. Inspired by the work done in Sekara et al. (2016) on fundamental structures in many-to-many synchronous interactions, we aim to provide insights into the fundamental structures on asynchronous many-to-many interactions - both regarding the temporal patterns and the aggregated networks. Thus, the the aim of this paper is to provide an initial addressing of the questions of whether there are any fundamental structures that are distinct for this grouping (many-to-many asynchronous) and whether there are any structures within the grouping that separates the networks.

1.1 Data description

We focus on data arising from the web forum Reddit, but highlight that this is a further subcategory of many-to-many asynchronous, where the users are anonymous and subscribe to content rather than people - an interesting and fertile hypothesis for a future study would be to investigate if there are fundamental differences between these networks and networks from social media with different characteristics.

* ✉ bskfr@dtu.dk

1.1.1 Terminologies

Throughout the paper the following terminologies are used frequently:

- A **subreddit** is a user-created area of interest where discussions on Reddit are organized.
- A Reddit **submission** can be text, link or a image submitted by a registered member. Submissions are made in a given subreddit and initialized a communication event.
- A **comment** is a response to the post that is active on Reddit. A comment can either be a direct response to the post or a response to any comment made on a post, thus creating a nested structure of a tree graph.
- The **author** is the user who is responsible for the submission and/or a comment in a communication event.

1.1.2 Data

As a starting point we investigates data from 5 different subreddits: *news*, *politics*, *football*, *pugs* and *puppies* in the period XXX.[1] These subreddits are chosen to have categories that are intuitively very different (eg. news vs. pugs), but also have categories that intuitively are very similar (eg. pugs and puppies).

1.2 Building the network

A communication event on Reddit is characterized with a submission, which in turn triggers an amount of activity in the form of comments. Thus, we can express this as a polytree, which we will refer to as the *submission-network*, $G_p = \langle V_p, E_p \rangle$, where each comment corresponds to a node $i \in V_p$ (with the submission as the root node $r \in V$), and the edges represents a reply for node j on node i, $(i, j) \in E_p$. We, further, generate a social network based on the implicit relationship between the authors of each node in the submission-network. We refer to this as the *author-network*, $G_a = \langle V_a, E_a \rangle$, where each comment corresponds to a node $i \in V_a$, and the edges represents a reply for node j on node i, $(i, j) \in E_a$. In Fig. 1 we show an example of the generation of an author-network, with the submission-network on the left and the corresponding author-network on the right[2]

2 Descriptive analysis

2.1 Temporal patterns

An aspect of the submission-network that we are interested in investigating is the temporal evolution of the communication events. This applies to the *lifetime* of the communication event, which we define as the time between the submission and the last comment,

[1] So far this period varies from each subreddit, but will likely be constant once we have a prove of concept idea.

[2] As a starting point, we work with an directed graph and discard all weights. However, this is subject to further considerations.

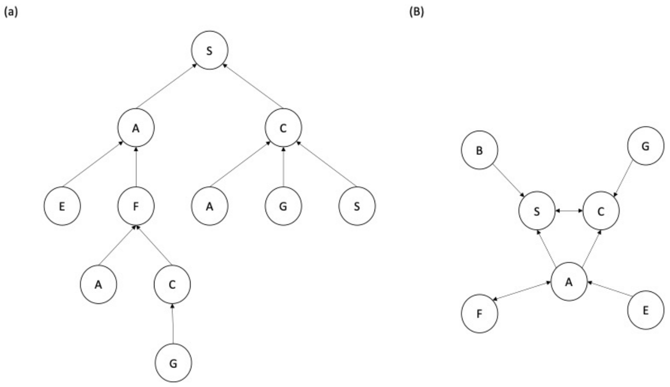


Fig. 1. Example of graph generation:

the *average/median time* of a comment (with respect to the submission) in a communication event, the lifetime of the communication event excluding the latest five percentile comments, which we refer to as *lifetime*$_{95}$. In Fig. 2 we illustrate the distribution of each variable for all retrieved submissions (top panel), as well as the distributions in each subreddit.

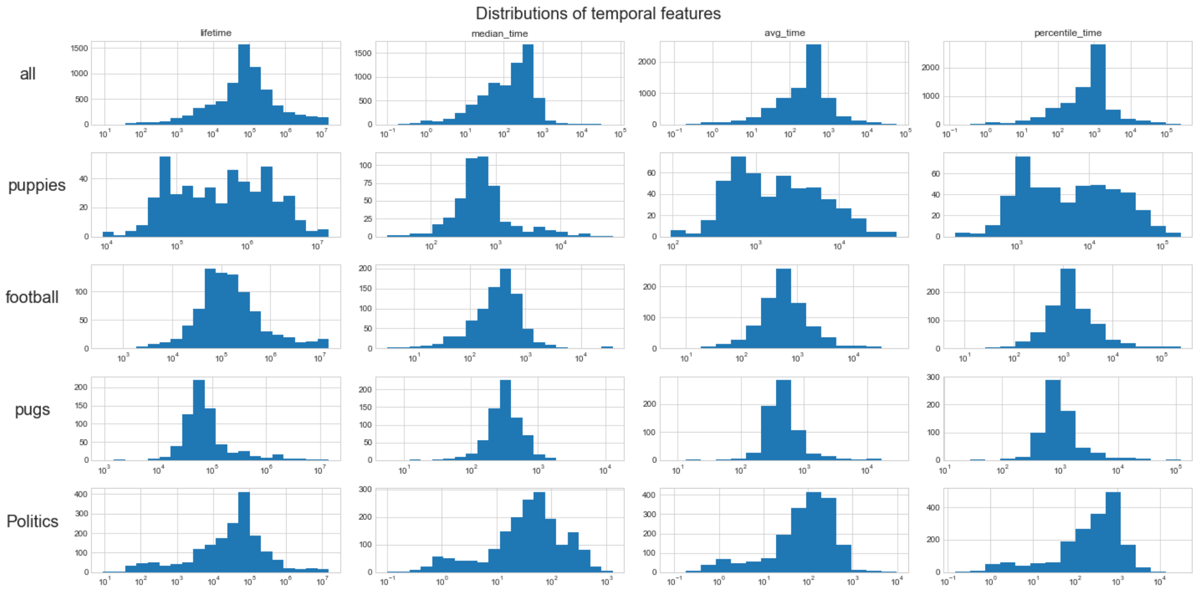


Fig. 2. Distributions of temporal features.

2.2 Submission specific characteristics

We next examine characteristics of each submission. Thus includes the total number of comments, the number of direct comments, i.e. the comments that are direct replies to the original submission and the max comment debth, as well as an adapted version of the H-index proposed in Gómez et al. (2008) and defined the following way: given a radial tree corresponding to a communication event and its comments organized in nesting levels, the H-index h of a submission is then the maximum nesting level i which has at least $h > i$ comments. In Fig. 7 we illustrate the distribution of each variable for all retrieved submissions (top panel), as well as the distributions in each subreddit.

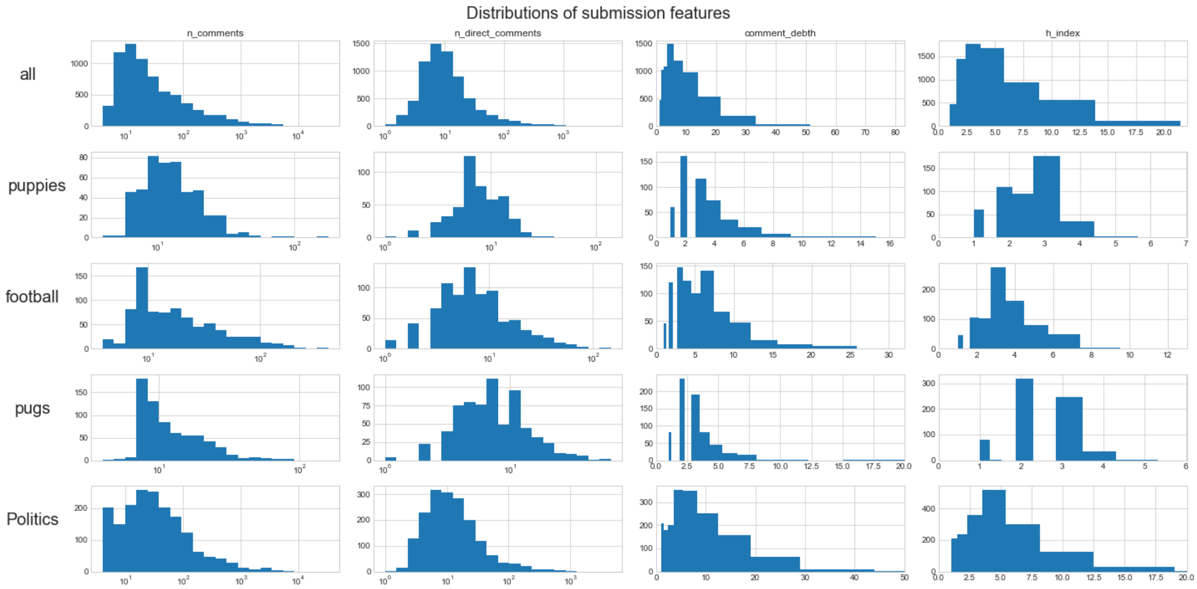


Fig. 3. Distributions of submission features.

2.3 The author-network

As a last for out descriptive analysis, we investigate the properties of the aggregated author-network. Here we initially look at the number of users present in each communication event, the number of edges (note that we are currently looking at an unweighted network) the size of the giant (weak) component, the recipricity coefficient, as well as the diameter and clustering coefficient (last two on an undirected version of the network). We illustrate the distributions in Fig. 4 and as above the top panel illustrates the distributions aggregated over subreddits, with the indivudual subreddits presented below.

3 Init ML

3.1 Methods

As an initial investigation we lastly build a simple random forrest model to help find patterns in the data. We use the subreddit as the target variable and thus mainly focus on the second question of whether there are any structures within the grouping that separates the communication events. We look at two models, where one is based on submission specific characteristics and the other the author specific characteristics (as presented above). For this we use 1,000 estimators and to measure the quality of the splits we use the gini impurity.

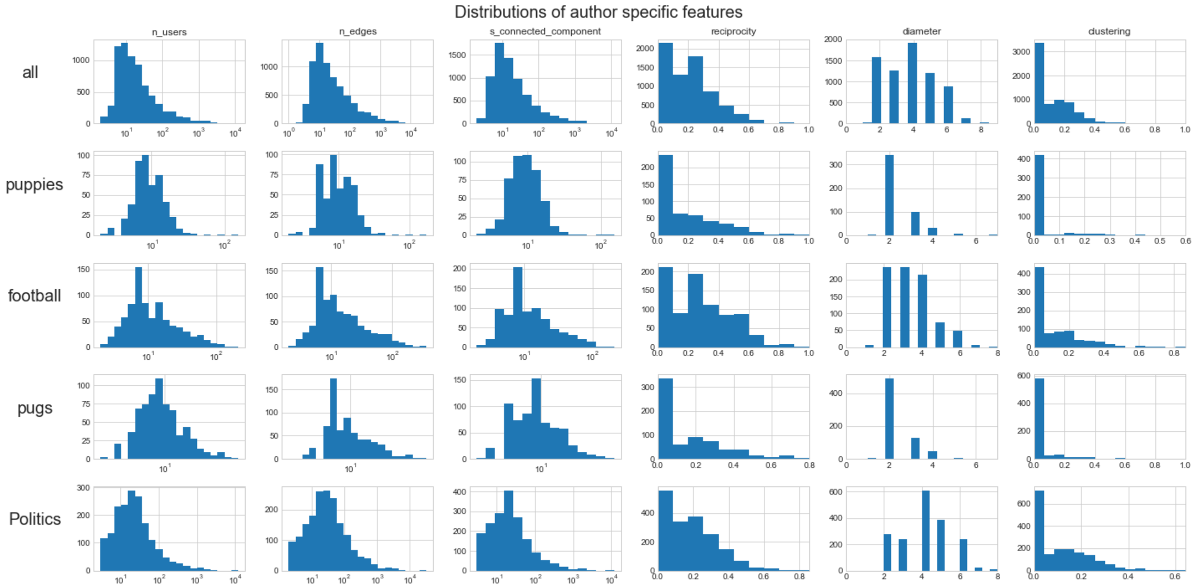


Fig. 4. Distributions of author features.

Balancing the data: We balance the data by random sampling according to the subreddit with the least submissions retrieved. For this PoC paper we have 482 submissions from each subreddit in out model.

Training and test data: We further split the data into a training and test set to provide an unbiased estimate of out models perfoamance - we use a 80/20 split.

Parameter configuration: To tune the parameters we use a simple gridsearch with six cross validations. We focus in the following parameters:

- *max_depth* which control the maximum and depth of each tree in the forrest. Large/deep trees introduce low bias and high variance.

- *max_features*, controlling the number of features sampled, i.e. the number of features to consider in each split. Trees with many features introduce low bias and high variance.

- *min_sample_split* the minimum number of samples required to split an internal node.

Data and code availability. All data and code used are available in the 'Funcdamental-structures-in-many-to-many-asynchronous-communication-networks' repository: `https://github.com/benj1003/Funcdamental-structures-in-many-to-many-asynchronous-communication-networks`.

3.2 Results

3.3 Post specific

xxx

3.4 author specific

For the author specific model we use only the features presented in Fig. 4. Via the gridsearch presented previously we get the optimal parameterset: *max_debth* = 5. *max_features*

$= 3$, *min_sample_split* $= 0.01$, with a CV score of 0.425 (note random guess is $1/5 = 0.2$).

Feature importance. As for the previous model the importance of each feature in the model is calculated using the CART algorithm calculated using the Scilearn module. We illustrate the cumulative importance of the features sorted by importance in the figure below.

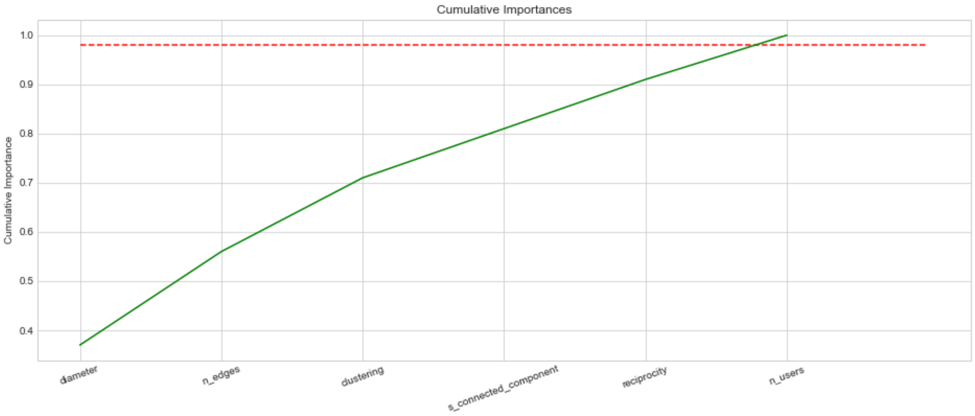


Fig. 5. Distributions of author features.

Test/training scores. We further investigates the presision and the accuracy of the model on both our training and test data. This show that the model doesn't do a great job predicting the correct subreddit. However, comparing against random guessing it does do significantly better.

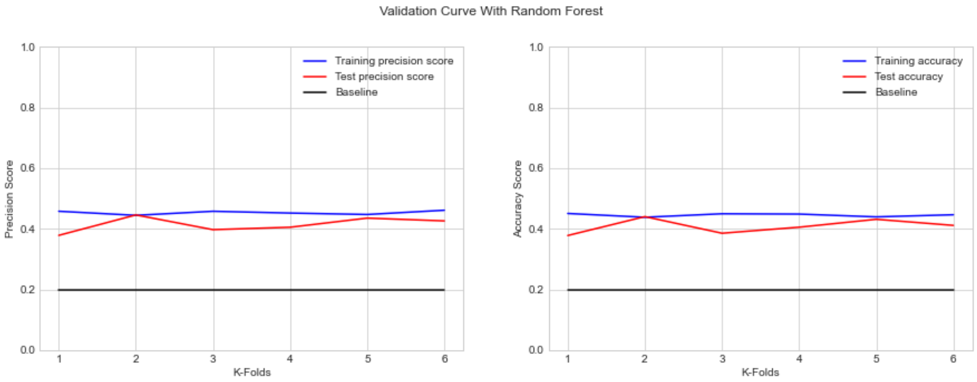
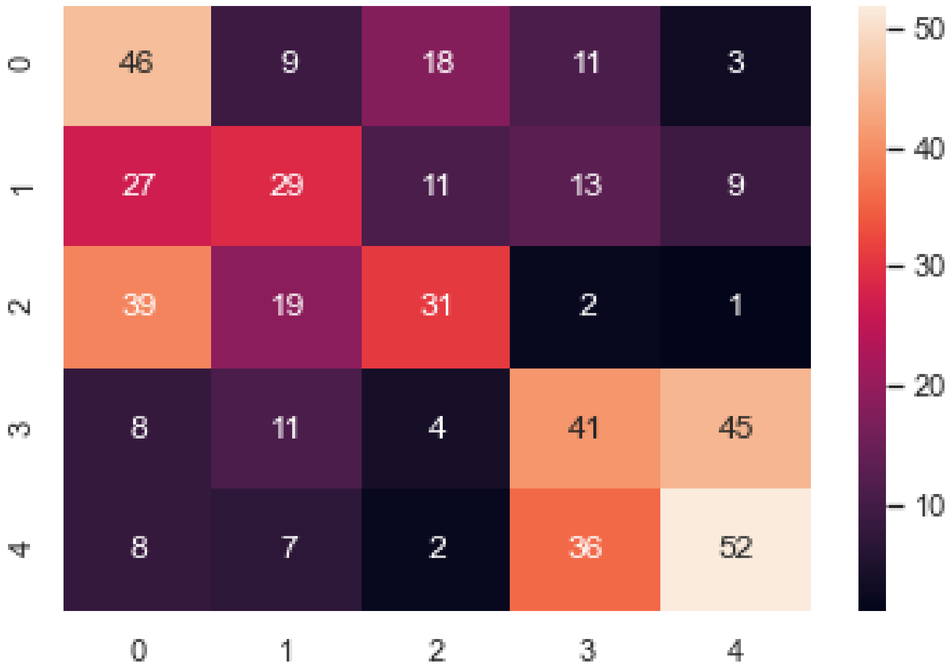


Fig. 6. Distributions of author features.

Confusion matrix. Lastly we look at the truth vs. predicted values in a confusion matrix to investigate if there are any trends in correct or wrong predictions. We see that there are some trends - eg. with 3 and 4, which corresponds to Puppies and Pugs, where the model very well predicts whether a submission is within one of these groups, but has a hard time seperating the two. Despite the relatively low accuracy/precision, this suggest that there are some clear patterns in the submissions that are different across different types subreddits. This, despite only considering the simplest possible aggregated author network.



	0	1	2	3	4
0	46	9	18	11	3
1	27	29	11	13	9
2	39	19	31	2	1
3	8	11	4	41	45
4	8	7	2	36	52

Fig. 7. Confusion matrix for author ML model.

References

Gómez, V., Kaltenbrunner, A., and López, V. (2008). Statistical analysis of the social network and discussion threads in slashdot. In Proceedings of the 17th International Conference on World Wide Web, WWW '08, page 645–654, New York, NY, USA. Association for Computing Machinery.

Lehmann, S. (2019). Fundamental structures in network temporal communication networks. Springer- Nature, New York. In Holme and Saramaki (Editors) Temporal Network Theory.

Sekara, V., Stopczynski, A., and Lehmann, S. (2016). Fundamental structures of dynamic social networks. Proceedings of the National Academy of Sciences, 113(36):9977–9982.

Thukral, S., Meisheri, H., Kataria, T., Agarwal, A., Verma, I., Chatterjee, A., and Dey, L. (2018). Analyzing behavioral trends in community driven discussion platforms like reddit. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 662–669.