Concentration Indices for Dialogue Dominance Phenomena in TV Series: The Case of the Big Bang Theory

Fronzetti Colladon, A., & Naldi, M.

Concentration indices for dialogue dominance phenomena in TV series: the case of the Big Bang Theory

Andrea Fronzetti Colladon and Maurizio Naldi

Abstract Dialogues in a TV series (especially in sitcoms) represent the main interaction among characters. Dialogues may exhibit concentration, with some characters dominating, or showing instead a choral action, where all characters contribute equally to the conversation. The degree of concentration represents a distinctive feature (a signature) of the TV series. In this paper we advocate the use of a concentration index (the Hirschman-Herfindahl Index) to examine dominance phenomena in TV series, and apply it to the Big Bang Theory TV series. The use of the concentration index allows us to reveal a declining trend in dialogue concentration as well as the decline of some characters and the emergence of others. We find the decline in dominance to be highly correlated with a decline in popularity. A stronger concentration is present for episodes (i.e., by analysing concentration of episodes rather than speaking lines), where the number of characters that dominate episodes is quite small.

1 Introduction

TV series are a steadily growing business, with the number of original scripted TV series in the U.S.A. exhibiting a CAGR (Compound Annual Growth Rate) of 11.1% from 2009 to 2017[1]. Writing their script is a delicate task, carried out with due care,

Andrea Fronzetti Colladon
University of Perugia, Department of Engineering, Via G. Duranti n. 93 06125 Perugia, Italy, e-mail: andrea.fronzetticolladon@unipg.it

Maurizio Naldi
LUMSA University, Dept. of Law, Economics, Politics and Modern languages, Via Marcantonio Colonna 19, 00192 Rome, Italy, e-mail: m.naldi@lumsa.it

[1] data from the Statista website https://www.statista.com/statistics/444870/scripted-primetime-tv-series-number-usa/

with the final aim of making the TV series successful [1]. A major task is the balance between characters and the prevalence given to some of them.

Analysing the relationship between characters through the tools of graph theory is a relatively recent area of research. The resulting networks are typically called character networks; a survey of the tools employed to automatically extract the character network is reported in [9], and the difficulties of the task are highlighted in [16] and [6]. Examples of the application of social network analysis to TV series are the analysis of the *Game of Thrones* performed in [2], and the analysis of narration of *Breaking Bad*, again *Game of Thrones*, and *House of Cards* in [4]. A two-mode network has been employed in [5] to analyse the dynamics of character activities and plots in movies. Social networks in fiction have also been employed to validate literary theories, e.g. to examine the relationship between the number of characters and the settings [8]. An even more ambitious example of trying to obtain the signature of a novel's story through the topological analysis of its character network is shown in [17].

In this paper, we wish to further explore the relationship among characters through the use of social network analysis tools by focussing on dialogues. In particular, we wish to identify dominant characters, i.e. characters that dominate dialogues. For that purpose we advocate the use of a concentration index borrowed from industrial economics, namely the Hirschman-Herfindahl Index (HHI).

We report the results obtained for the Big Bang Theory (BBT) series. We show that:

- a steady declining trend is present for dominance in the number of speaking lines;
- the rank-size distribution of speaking lines among the major characters is roughly linear;
- a pattern is also present in the characters dominating the scene, with some declining while others emerge;
- episodes are dominated by a very small number of characters.

The paper is organized as follows. We describe our Big Bang Theory dataset in Section 2 and the associated graph in Section 3. In Section 4, we introduce the concentration (dominance) index and show the results of its application to the BBT series.

2 The dataset

Our analysis is based on the set of scripts for the Big Bang Theory (BBT) TV series. In this section we describe the main characteristics of that dataset, which has also been employed in [7].

The Big Bang Theory is an American television sitcom, premiered on CBS in 2007. It has now reached its twelfth season. After a slow start (ranking 68th in the first season and 40th in its second one), it ranked as CBS's highest-rated show in that evening on its first episode in the third season.

The dialogues have been retrieved from the BBT transcripts site[2]. The script reports the dialogues as a sequence of speaking lines, each line being formed of the speaking character name and the text of his/her speech (an uninterrupted text by a character counts as one speaking line, irrespective of the actual length of the text; a speaking line ends when another character takes turns). An excerpt of the dialogues is shown hereafter.

LEONARD : I'm sure she'll still love him.
SHELDON : I wouldn't.
LEONARD : Well, what do you want to do?
SHELDON : I want to leave.

We have examined all the episodes from the initial one of Season 1 to the 24th episode of Season 9, for a total of 207 episodes.

3 Representation of dialogues

Our aim is to detect the presence of dominant characters in the series. We will focus on dialogues, since the series is essentially a sitcom (see [14] for an introduction to the genre, and [10] for the collocation of the Big Bang Theory within that genre) and is therefore based on dialogues (this is different, e.g., from what happens in an action movie, where the interaction is mainly physical). The notion of dominance is associated to those characters speaking most of the time. In this section, we see how we can extract the interaction structure associated to dialogues for the purpose of detecting dominance phenomena.

As reported in Section 2, the dialogues are shown in the script as sequences of speaking lines, each line representing a talkspurt by a single character. Each character interacts with the following character taking turns. We can represent that interaction through a graph, which describes the social network embedded in the dialogues. Our dialogue-based approach is different from what is done in [2], where any kind of interaction is included, or in [3], where two characters are connected if their names occur a certain number of words apart from each other. However, for the time being, we neglect the actual content of the dialogues or the sentiments conveyed by them.

The graph is built by connecting two characters if they talk to each other. Actually, we build a weighted directed graph. The nodes represent the characters; there is a link from character A to character B if A speaks to B (we infer that A speaks to B when a speaking line by A is followed by a speaking line by B); the weight of the link represents the number of times when A speaks to B.

As an example, the resulting dialogue graph is shown in Fig. 1 for Episode 1 of the first season. In order not to garble the graph, we have not reported the links concerning single instances, i.e. when character A speaks just once to character B.

As we can see, just two links are two-way, which means that just two talking relationships are reciprocated (namely those between Sheldon and Leonard, and

[2] https://bigbangtrans.wordpress.com

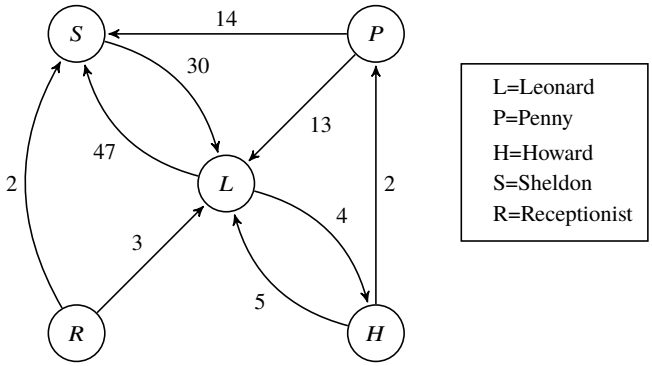

Fig. 1 Dialogue graph for Episode 1 of Season 1

between Howard and Leonard). For this example, we can compute the density (the ratio of actual links to the maximum potential number of links for a graph of that size), which is just 0.45.

An alternative way to describe the interaction between the characters is the associated (weighted) adjacency matrix D, which is reported below, where we have employed a mapping between character names and indices as suggested by the order in the legend in Fig. 1 (i.e., 1 for Leonard, 2 for Penny, and so on). The element d_{ij} of D is the number of times A speaks to B.

$$\mathbf{D} = \begin{pmatrix} 0 & 0 & 4 & 47 & 0 \\ 13 & 0 & 0 & 14 & 0 \\ 5 & 2 & 0 & 0 & 0 \\ 30 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 2 & 0 \end{pmatrix}$$

In the graph, the out-degree of each node represents the number of times that the character speaks before somebody, i.e. its number of speaking lines in the script. Again for the first episode, we show the speaking lines for the group of 5 characters in the following vector, which are, otherwise stated, the out-degrees of each node, or the row sums of the adjacency matrix.

$$\begin{pmatrix} 51 \\ 27 \\ 7 \\ 30 \\ 5 \end{pmatrix}$$

4 Dominance in dialogues

After building the dialogue graph and explained the notion of dominance in this context, we wish to analyse if such dominance phenomena are present, and to what extent, in The Big Bang Theory series. For that purpose, in this section we introduce a dominance index and apply it to the BBT scripts. In addition, we identify the dominant characters throughout the series.

As a dominance index, we borrow a concentration index from the field of industrial economics, named the Hirschman-Herfindahl Index (or HHI, for short) [13, 11]. In order to see if an industry is concentrated in the hands of few companies (i.e., if a few companies dominate the market), the HHI was defined as the sum of the squared market shares of all the companies in the market: the higher the HHI, the more concentrated the market (i.e., dominated by a few firms). In our case, we similarly define HHI as our dominance index by considering the number of times a character speaks to any other character (i.e., the number of speaking lines of that character, in the theatre jargon). The equivalent of the *market share* in this context is therefore the fraction of speaking lines of each character with respect to the overall number of speaking lines in the episode. On the graph, it is the out-degree of that character normalized by the sum of all the out-degrees.

By recalling the definition of the weighted adjacency matrix **D**, for an episode in which n characters appear, the HHI is

$$\mathrm{HHI} = \frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{n} d_{ij} \right)^2}{\left(\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} \right)^2} \tag{1}$$

The HHI takes values in the $[1/n, 1]$ range, with $1/n$ representing perfect equipartition of the market (in our case, perfect equidistribution of speaking lines among all the characters), and 1 representing a monopoly (in our case, a monologue by one character). As to intermediate cases, in order to determine whether an HHI value denotes a strong dominance by one or more characters, we can adopt the classification suggested in [15]: a value lower than 0.15 denotes unconcentrated markets, while a value larger than 0.25 denotes highly concentrated markets, and intermediate values represent moderately concentrated markets. The HHI has already been used to analyze dominance in dialogues in [7] for TV series and [12] for personal finance forums.

We report an example of computation of the HHI for the graph of Fig. 1 in Table 1.

We can now compute the HHI for each episode in the series. In Fig. 2, we report the average HHI across each season. We can now see that the concentration in dialogues has been steadily decreasing over the years. The dominance index fell into the moderate concentration region as early as Season 3, and bordered on the unconcentrated region starting from Season 7. We do not know whether this was a deliberate choice of the series authors, but we note that the series has moved from

Character	Out-Degree	Dialogue share	Squared dialogue share
Leonard	51	0.425	0.1806
Sheldon	30	0.25	0.0625
Penny	27	0.225	0.0506
Howard	7	0.058	0.0034
Receptionist	5	0.042	0.0018
HHI			0.2989

Table 1 HHI computation for the graph of Fig. 1

episodes with a few dominant characters to a more choral action. At the same time, we observe a moderate dispersion around the average value, accounting for roughly ±20% of the average.

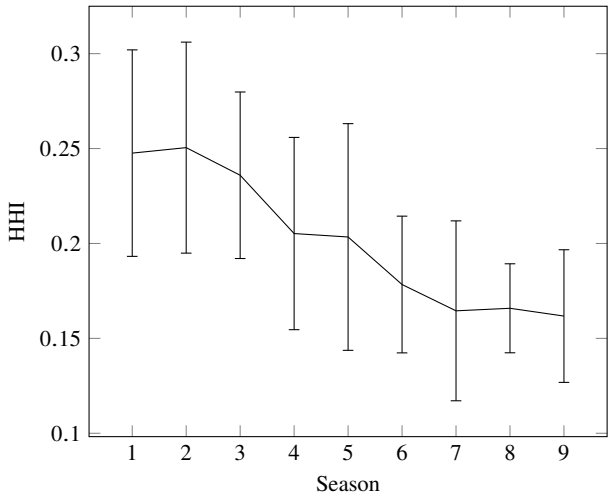

Fig. 2 Time evolution of HHI

It is interesting to compare that trend in dominance with a similar trend observed for viewers' ratings collected on the IMDB (Internet Movie Database) platform, reported in Fig. 3 [7]. The resulting correlation coefficient is 0.847, showing that the two trends are quite similar. Though we cannot state a causal relationship, we observe that the declining popularity has been associated to the shift from a dominance situation to a more choral action.

So far, we have considered the presence of dominance phenomena without specifying who's dominating the dialogues.

We consider first the dominance throughout the series by analysing the overall number of speaking lines spoken by each character. In Fig. 4, we report that number for the major characters. We see that we are far from an equidistribution even among

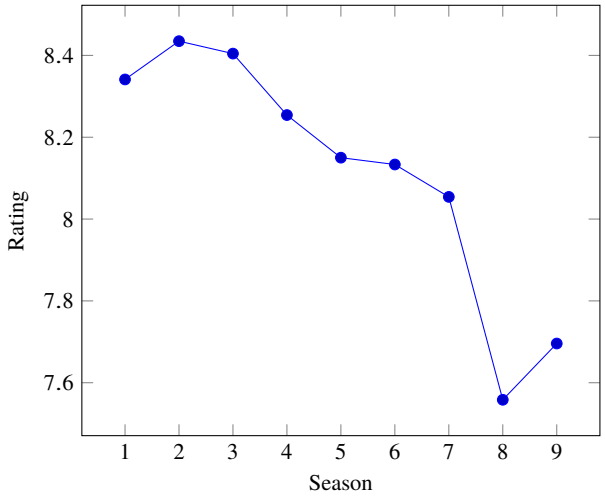

Fig. 3 Viewers' ratings

the major characters. The character speaking most over the series is Sheldon, followed by Leonard and Penny. The decay appears to be quite linear in the rank.

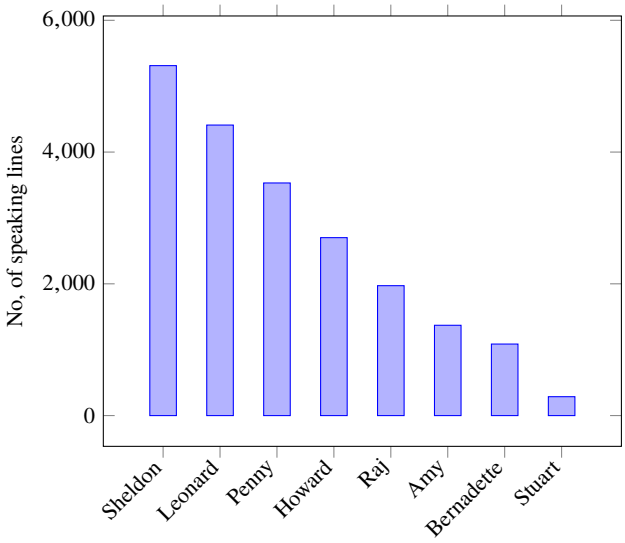

Fig. 4 Number of speaking lines per character (whole series)

Delving deeper, we may wish to see if the dominant character was the same throughout the series, or an alteration of dominant characters was present, or some characters faded along the series to give room to emerging characters. We can assess

that by identifying the dominant character in each episode. In Fig. 5, we report the number of episodes where each character was dominant. Actually, we see several patterns here. After building up in Seasons 1 to 4, Sheldon lost ground in Season 5 to regain his leading role afterwards. Instead Leonard following a rather steady decline throughout the seasons. Penny gradually gained prominence, but reached her peak in Season 7 and fell somewhat during the last two seasons.

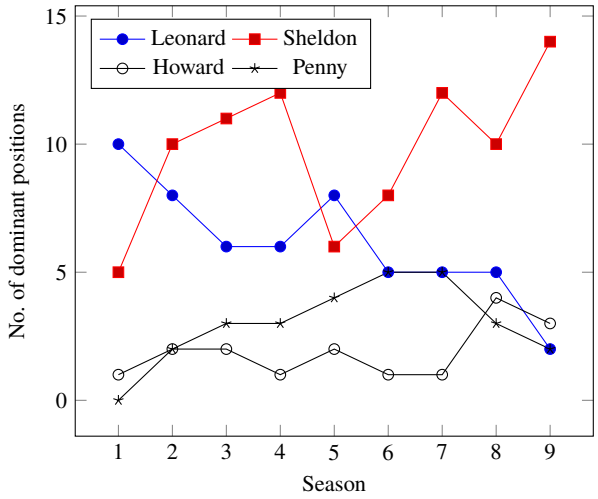

Fig. 5 Dynamics of dominating characters

The cumulative plot of speaking lines in Fig. 6 shows the gradual takeover by Sheldon, who became the overall leading character as early as in Season 3.

Since we have now moved to considering episodes rather than speaking lines as the measure of dominance, it is interesting to examine if the concentration metric is different from what we saw for the number of speaking lines. If we apply the definition of the HHI to the number of episodes where a character appears as dominant, we obtain the graph in Fig. 7. If we compare this graph with that of Fig. 2, we get a bit different story. First, the figures are quite higher: the concentration is much larger when we measure it over the number of dominated episodes rather than over the number of speaking lines, which means that episodes are assigned to a small number of dominating characters. Secondly, we do not observe the steady declining trend as in Fig. 2. Rather we have a stronger concentration in the first and the last season, but in between we see a concentration fluctuating around a central value (roughly 0.35).

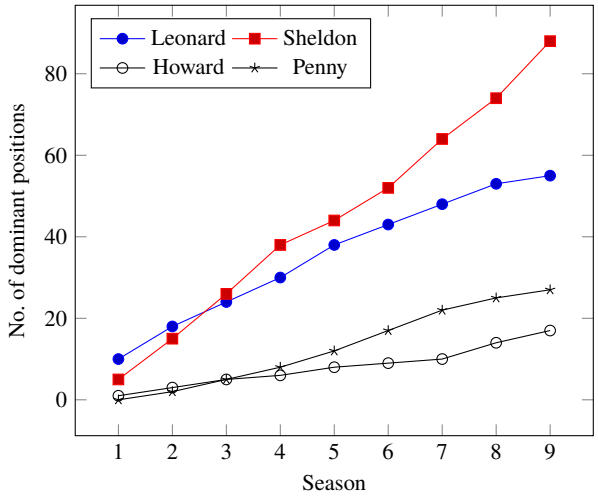

Fig. 6 Cumulative dynamics of dominating characters

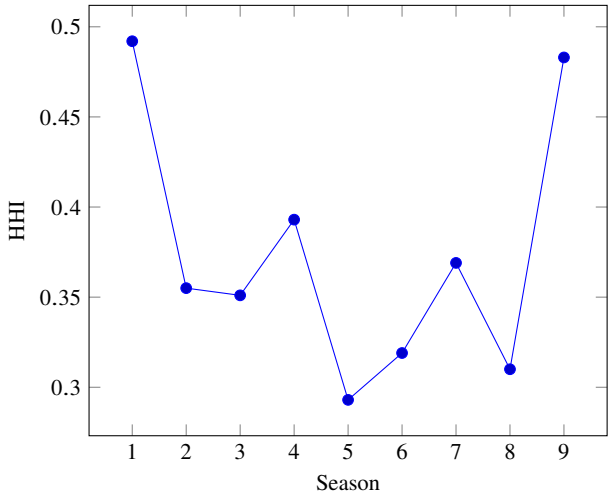

Fig. 7 HHI dynamics for the number of episodes

5 Conclusions

We have introduced the use of a concentration index, namely the Hirschman-Herfindahl Index, borrowed from industrial economics, to analyse dominance phenomena in the dialogues of TV series. We have performed a concentration analysis of the Big Bang Theory TV series. The analysis allows to reveal a declining trend in the concentration of dialogues over the years, i.e. the passage from a few dominating characters to a more choral action. However, the number of characters that dominate

episodes is rather small. The distribution of speaking lines among characters over the whole series exhibits a linear rank-size relationship. However, the analysis of dominance by specific characters over seasons allows to detect a pattern where some characters decline in importance and others emerge.

The adoption of a dominance index is therefore a valuable aid to detect a relevant stylistic feature of a TV series such as the dominance of some characters in dialogues, and to investigate the association between stylistic features of the series and its performance. A refinement of the analysis can be envisaged where the recipient of each speaking line is more accurately identified.

Acknowledgements Maurizio Naldi has been partially supported by the Italian Ministry of Education, University, and Research (MIUR) under the national research projects PRIN AHeAd #20174LF3T8.

References

1. Allrath, G., Gymnich, M., Surkamp, C.: Introduction: Towards a narratology of tv series. In: Narrative strategies in television series, pp. 1–43. Springer (2005)
2. Beveridge, A., Shan, J.: Network of thrones. Math Horizons **23**(4), 18–22 (2016)
3. Bonato, A., D'Angelo, D.R., Elenberg, E.R., Gleich, D.F., Hou, Y.: Mining and modeling character networks. In: Algorithms and Models for the Web Graph: 13th International Workshop, WAW 2016, Montreal, QC, Canada, December 14–15, 2016, Proceedings 13, pp. 100–114. Springer (2016)
4. Bost, X., Labatut, V., Gueye, S., Linarès, G.: Narrative smoothing: dynamic conversational network for the analysis of tv series plots. In: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1111–1118. IEEE Press (2016)
5. Chao, D., Kanno, T., Furuta, K., Lin, C.: Representing stories as interdependent dynamics of character activities and plots: A two-mode network relational event model. Digital Scholarship in the Humanities (2018)
6. Edwards, M., Mitchell, L., Tuke, J., Roughan, M.: The one comparing narrative social network extraction techniques. arXiv preprint arXiv:1811.01467 (2018)
7. Fronzetti Colladon, A., Naldi, M.: Predicting the performance of tv series through textual and network analysis: The case of big bang theory. PloS one **14**(11) (2019)
8. Jayannavar, P., Agarwal, A., Ju, M., Rambow, O.: Validating literary theories using automatic social network extraction. In: Proceedings of the Fourth Workshop on Computational Linguistics for Literature, pp. 32–41 (2015)
9. Labatut, V., Bost, X.: Extraction and analysis of fictional character networks: A survey. ACM Computing Surveys (CSUR) **52**(5), 89 (2019)
10. Ma, Z., Jiang, M.: Interpretation of verbal humor in the sitcom the big bang theory from the perspective of adaptation-relevance theory. Theory & Practice in Language Studies **3**(12) (2013)
11. Naldi, M.: Concentration indices and Zipf's law. Economics Letters **78**(3), 329–334 (2003)
12. Naldi, M.: Interactions and sentiment in personal finance forums: An exploratory analysis. Information **10**(7), 237 (2019)
13. Rhoades, S.A.: The Herfindahl-Hirschman Index. Fed. Res. Bull. **79**, 188 (1993)
14. Smith, E.S.: Writing television sitcoms. Penguin (1999)
15. The U.S. Department of Justice and the Federal Trade Commission: Horizontal Merger Guidelines (19 August 2010)

16. Vala, H., Jurgens, D., Piper, A., Ruths, D.: Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 769–774 (2015)
17. Waumans, M.C., Nicodème, T., Bersini, H.: Topology analysis of social networks extracted from literature. PloS one **10**(6), e0126470 (2015)