Modeling How Data Journalism Shapes Discussion Networks

Avner Kantor[1,2]; Sheizaf Rafaeli[3,4]
[1]University of Haifa; [2]Cental Bureau of Statistics; [3]Shenkar College; [4]Samuel Neaman Institute;

Abstract

This study investigates how information accessibility in data journalism—via statistical information, sources, and visualizations—shapes the structure of online discussion networks. Based on *4,873 New York Times* reply networks, we assess direct and mediated effects on network size, diameter, and transitivity. Mediation analysis reveals that static visualizations and statistical information significantly influence these structural properties, suggesting that accessible information plays a key role in shaping online discourse.

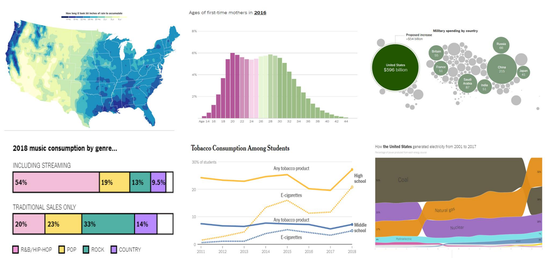

Figure 1. Data visualizations elements from stories in the data journalism section *The* Upshot of *The New York Times.*

Background and Framework

As digital media evolves, online comment sections have become key spaces for public deliberation. Data journalism (DJ) has emerged as a genre that integrates statistics, sources, and visualizations into storytelling to enhance clarity, credibility, and accessibility.

We conceptualize discussions as social networks, where users are nodes and replies are directed links—revealing patterns of interaction and audience engagement.

Research Question: *How does data journalism affect discussion network structure?*

Using a mediation framework (Figure 2), we assess the direct and indirect effects of DJ on three structural measures:

- **Network size**: the number of unique participants,
- **Directed diameter**: the longest shortest path between participants,
- **Transitivity**: the likelihood that participants form tightly connected clusters.

We examine the role of three mediators: **statistical information**, **information sources**, and **data visualizations**.

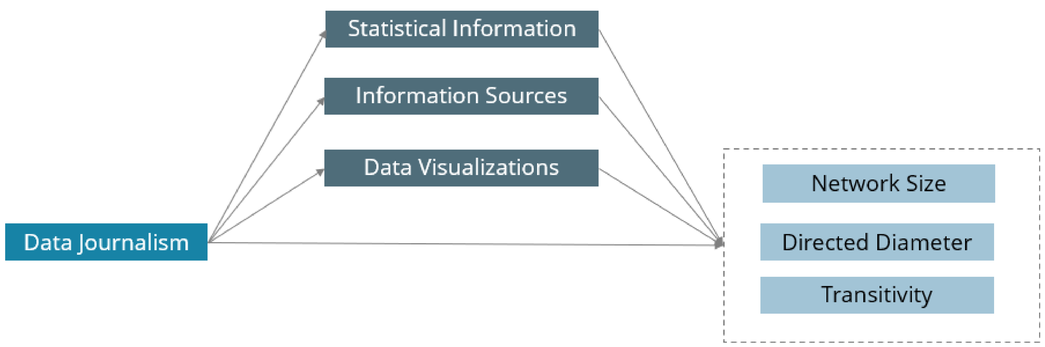

Figure 2. Conceptual model illustrating the direct and mediated effects of data journalism on discussion network.

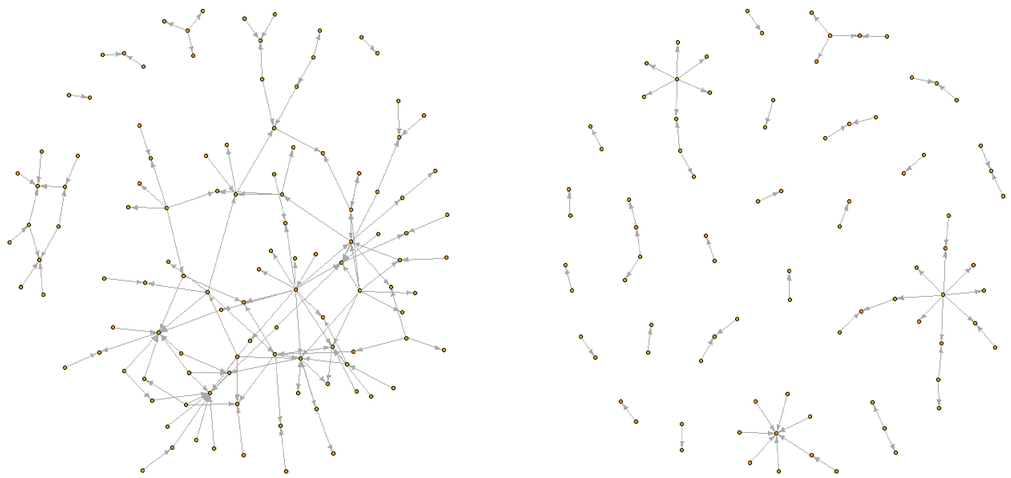

Figure 3. Example discussion networks from the dataset: (A) data journalism and (B) traditional journalism. These networks illustrate typical differences in size, diameter, and clustering. Mediating variables are not depicted.

A: 114 nodes, diameter = 5, transitivity = 0.032

B: 85 nodes, diameter = 2, transitivity = 0.000

Methodology

We collected user comment data from *The New York Times*. Using the NYT API, we retrieved stories and their associated comment sections from *The Upshot* —its DJ section—and from other sections, spanning the years 2014 to 2022. The final dataset includes 4,873 stories and ~521K user comments.

Story features, including statistical information, sources, and static visualizations, were extracted through HTML and CSS parsing. Discussion networks were constructed as reply networks, in which nodes represent commenters and directed edges indicate reply actions. These form directed and weighted graphs, with degree distributions that follow a power-law—consistent with social network dynamics.

Network measures were calculated at the story level and analyzed using Hayes' PROCESS macro (Model 4) to assess mediation effects.

Conclusion

As shown in the example networks in Figure 3, the results reveal clear differences in discussion network structure.

DJ is associated with **larger discussion networks**, an effect that is mediated by statistical information, information sources, and static visualizations. These larger networks suggest broader participation and greater exposure to diverse perspectives.

DJ is also associated with **a greater directed diameter**, mediated by information sources, indicating more dispersed and wide-reaching discussions.

DJ further **increases transitivity**, mediated by static visualizations, reflecting stronger local clustering and more tightly knit interactions.