Origraph: Interactive Network Wrangling

Supplementary Material

Alex Bigelow
abigelow@sci.utah.edu
University of Utah

Carolina Nobre
cnobre@sci.utah.edu
University of Utah

Miriah Meyer
miriah@cs.utah.edu
University of Utah

Alexander Lex
alex@sci.utah.edu
University of Utah

1 Comparison to Other Tools

Modeling / Reshaping	Origraph	Ploceus	Orion	Ploceus Term	Orion Term
Connect/Disconnect Nodes	yes	yes	yes	Create Connection	Link
Promote Attribute	yes	yes	yes	Create	Promote
Facet Node/Edge Class	yes	yes	yes	Slice n'Dice	Split
Convert Between Node/Edge	yes	no	no		
Project Edges	yes	partially*	partially*		
Create Supernodes	yes	no	no		
Roll Up Edges	yes	yes	yes		
Item Operations					
Filter by Attributes	yes	no†	yes		
Connectivity-Based Filtering	yes	no	no		
Attribute Operations					
Change Edge Direction	yes	no	no		
Derive In-Class Attributes	yes	yes	yes		
Connectivity-Based Attribute Derivation	yes	no‡	no		

Table 1: Table of operations supported by Origraph, Orion [2], and Ploceus [4]. In some cases, operations are only partially supported, as denoted by the superscripts.
* Only for immediate neighbors
† Ploceus can identify specific node values through the search function
‡ Ploceus does allow for duplication of attributes across tables

2 Gender Bias in Movies

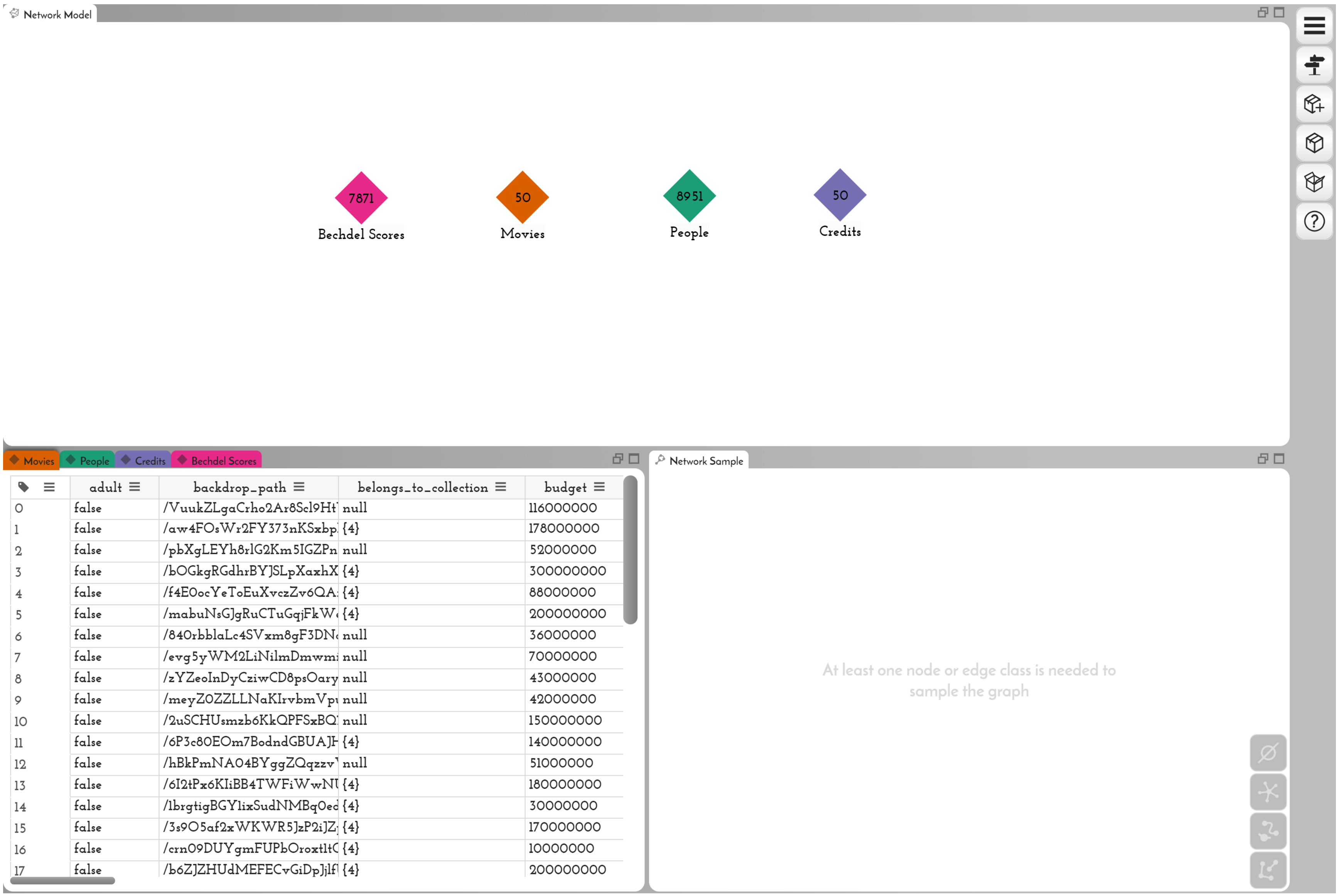

Figure M1: In this use case, described in Section 9.1 in the paper, we investigate gender bias in movies. Here we laded raw movies datasets, consisting of TMDB Movies, Credits, and People [1] as well as Bechdel Scores [6].

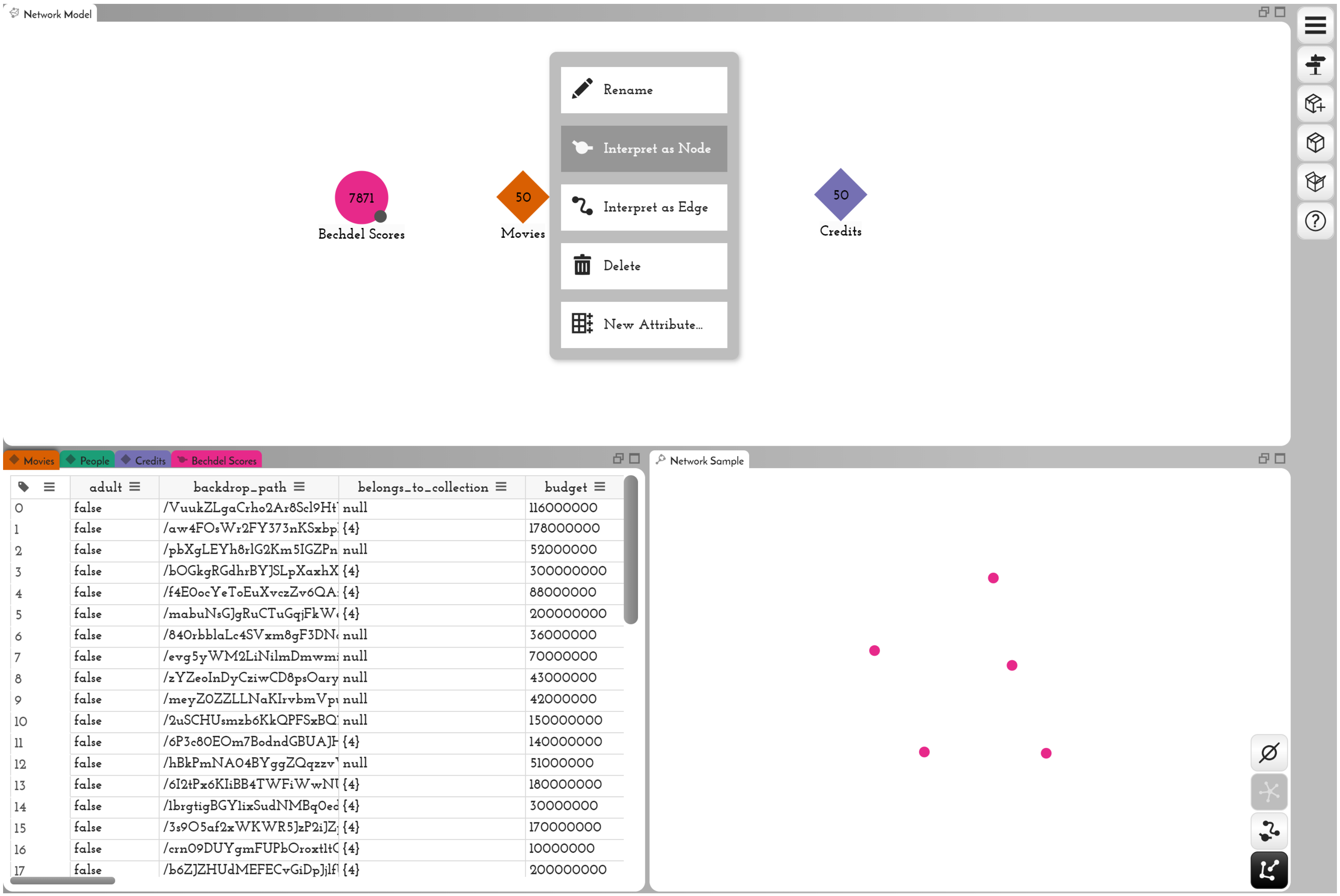

Figure M2: The Bechdel Scores and Movies tables are from different sources, but both refer to movies—in order to combine them, we first convert each to node classes.

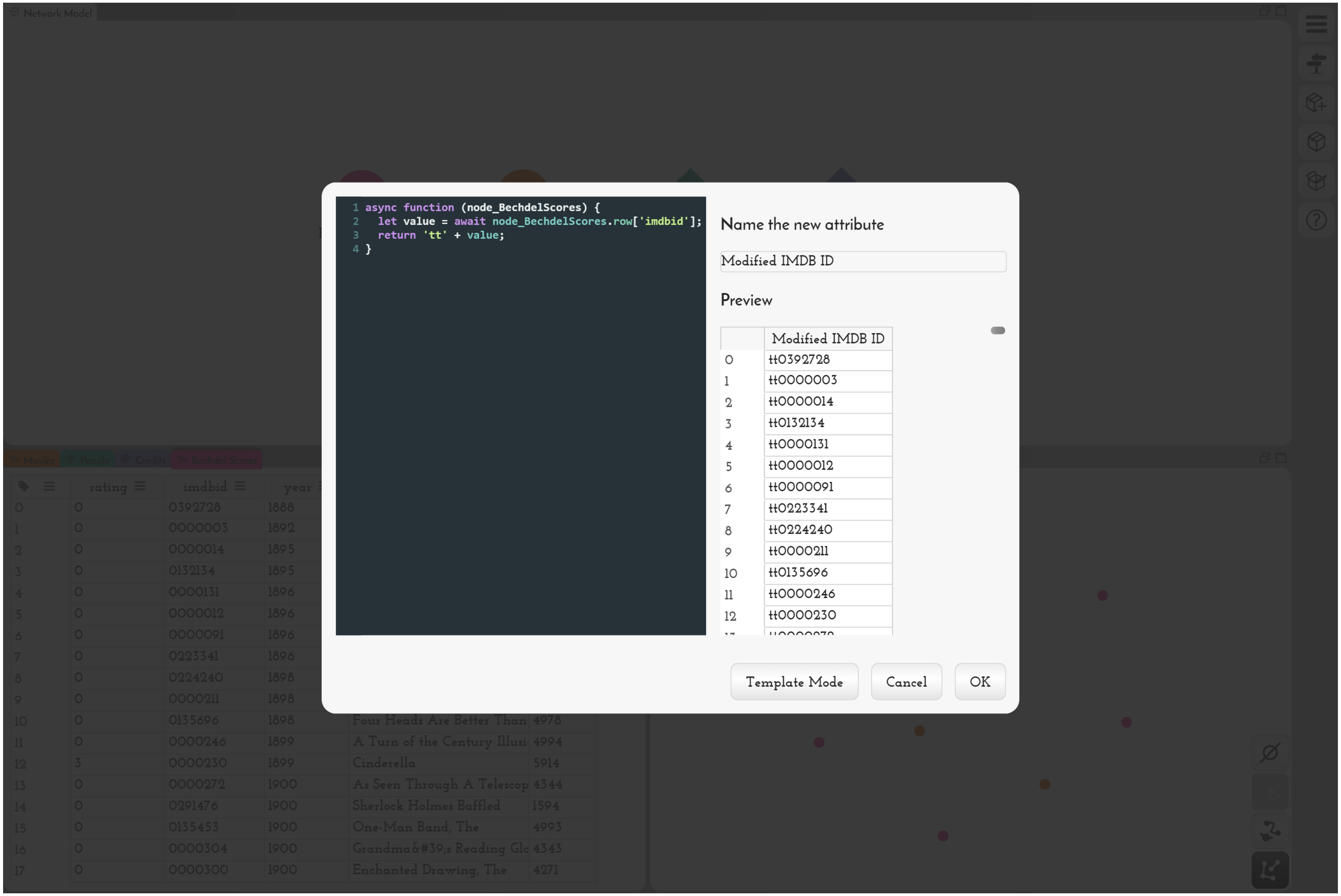

Figure M3: We want to connect both classes through IMDB IDs; however, the Movies table has a "tt" string prepended to its `imdb_id` attribute. Here, we derive a new attribute based on the Bechdel Scores' IMDB IDs, to make connection possible.

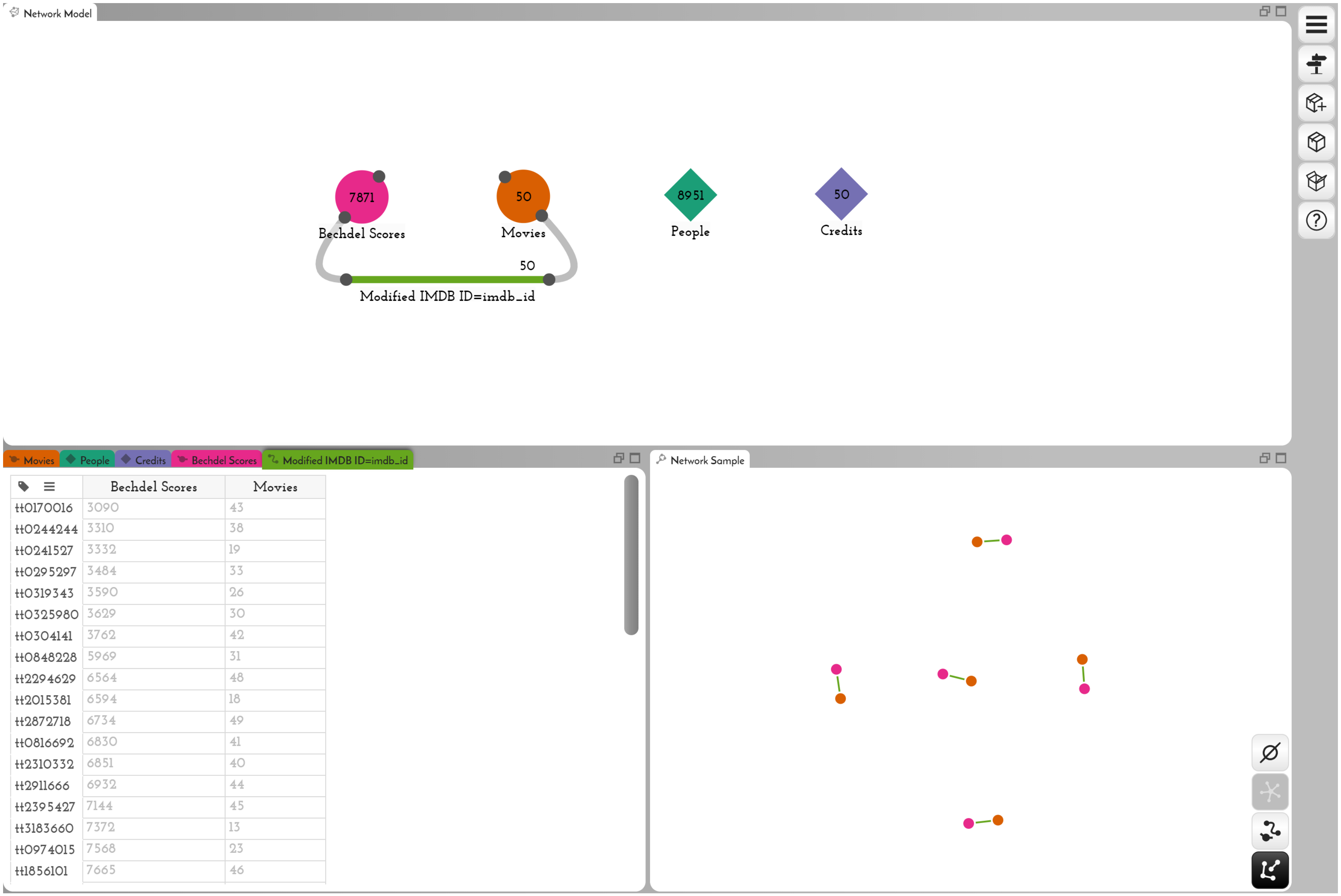

Figure M4: The state of the network model after connecting Bechdel Scores and Movies

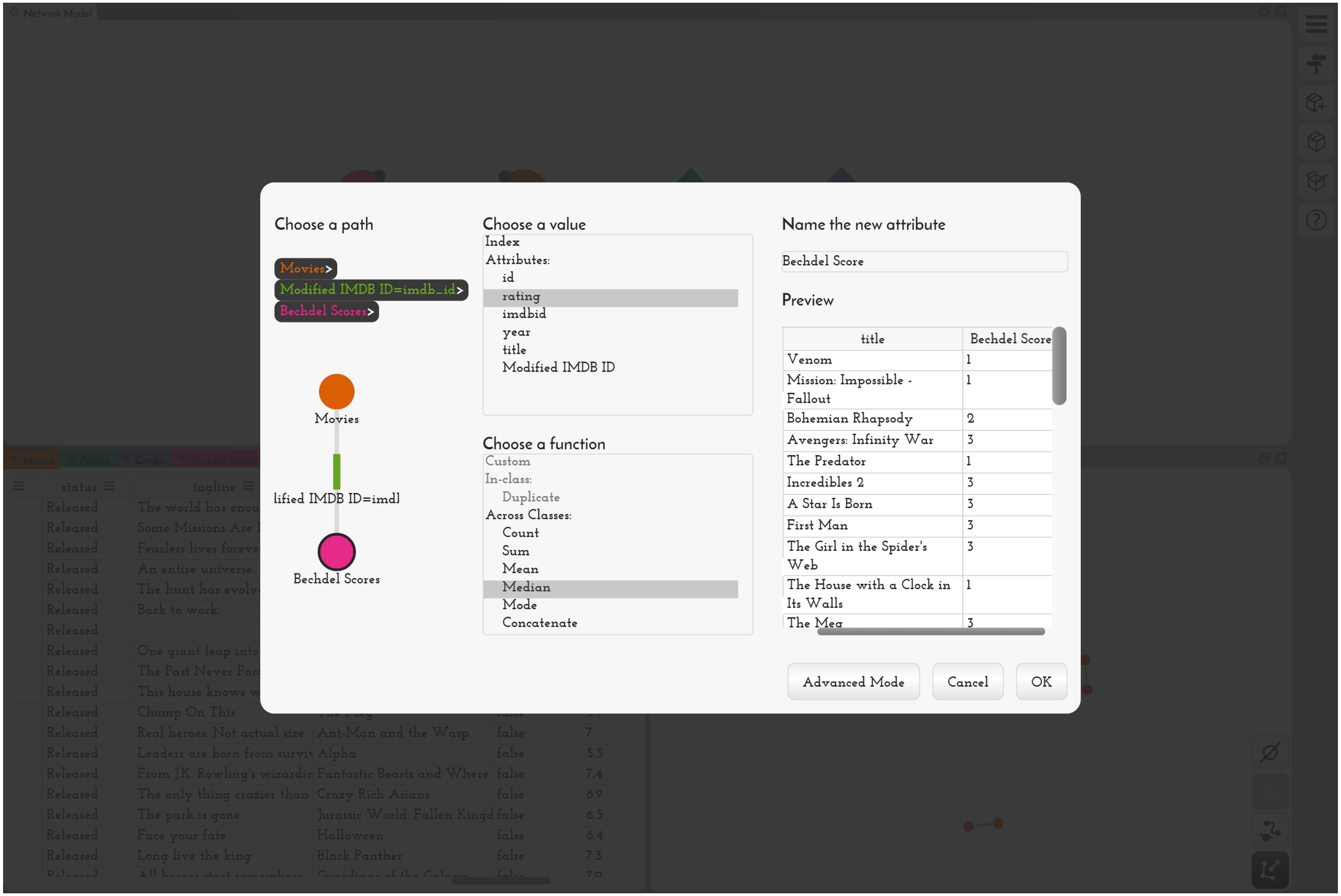

Figure M5: The standard attribute derivation interface, showing how attributes can be derived from connected nodes and edges. Here, we are computing the median rating for a Movie, based on its connected Bechdel Score. Because Movies and Bechdel Scores have a 1-1 mapping, this de-facto simply copies the rating.

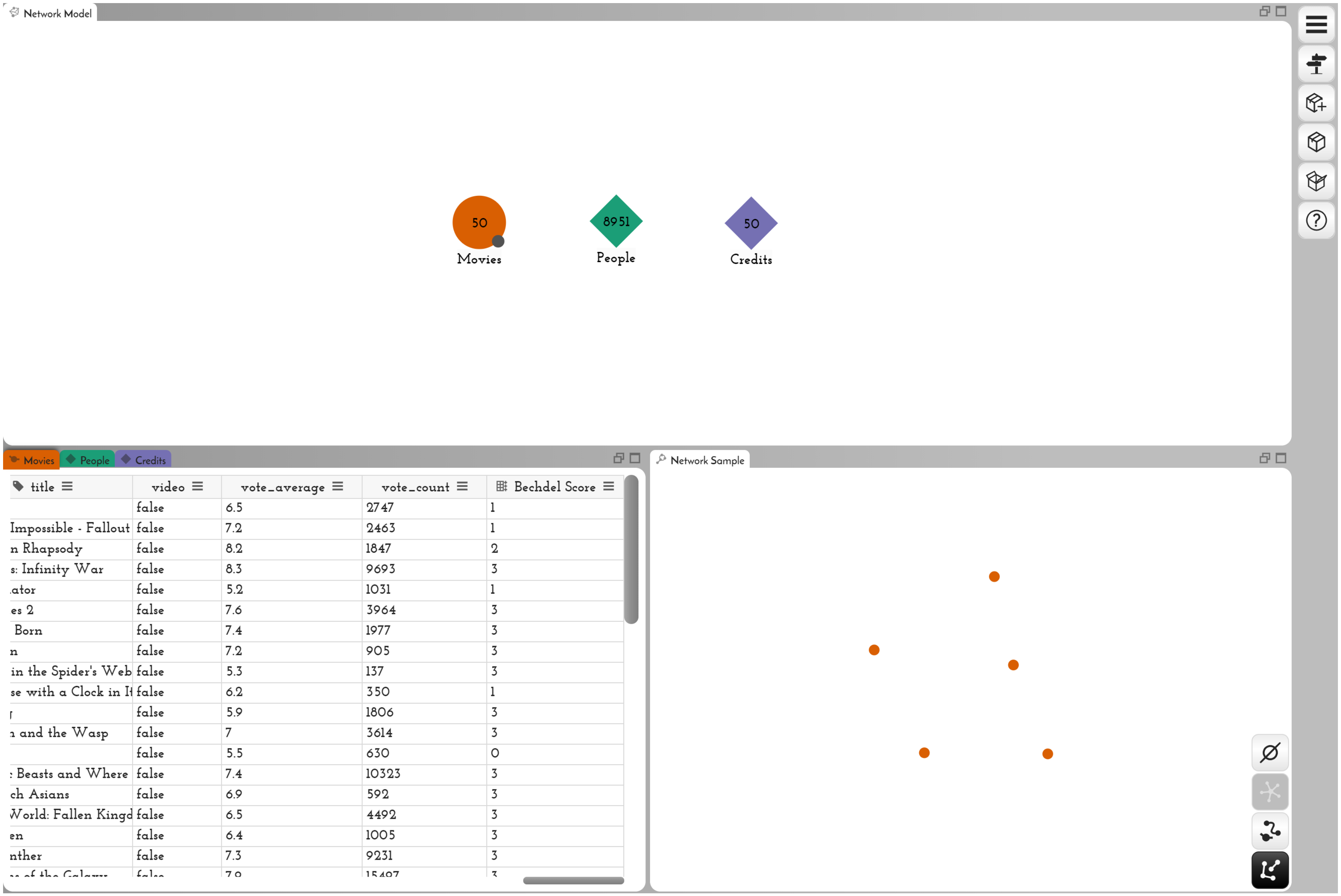

Figure M6: The remaining classes, now that the Bechdel Scores class is no longer needed, and has been deleted.

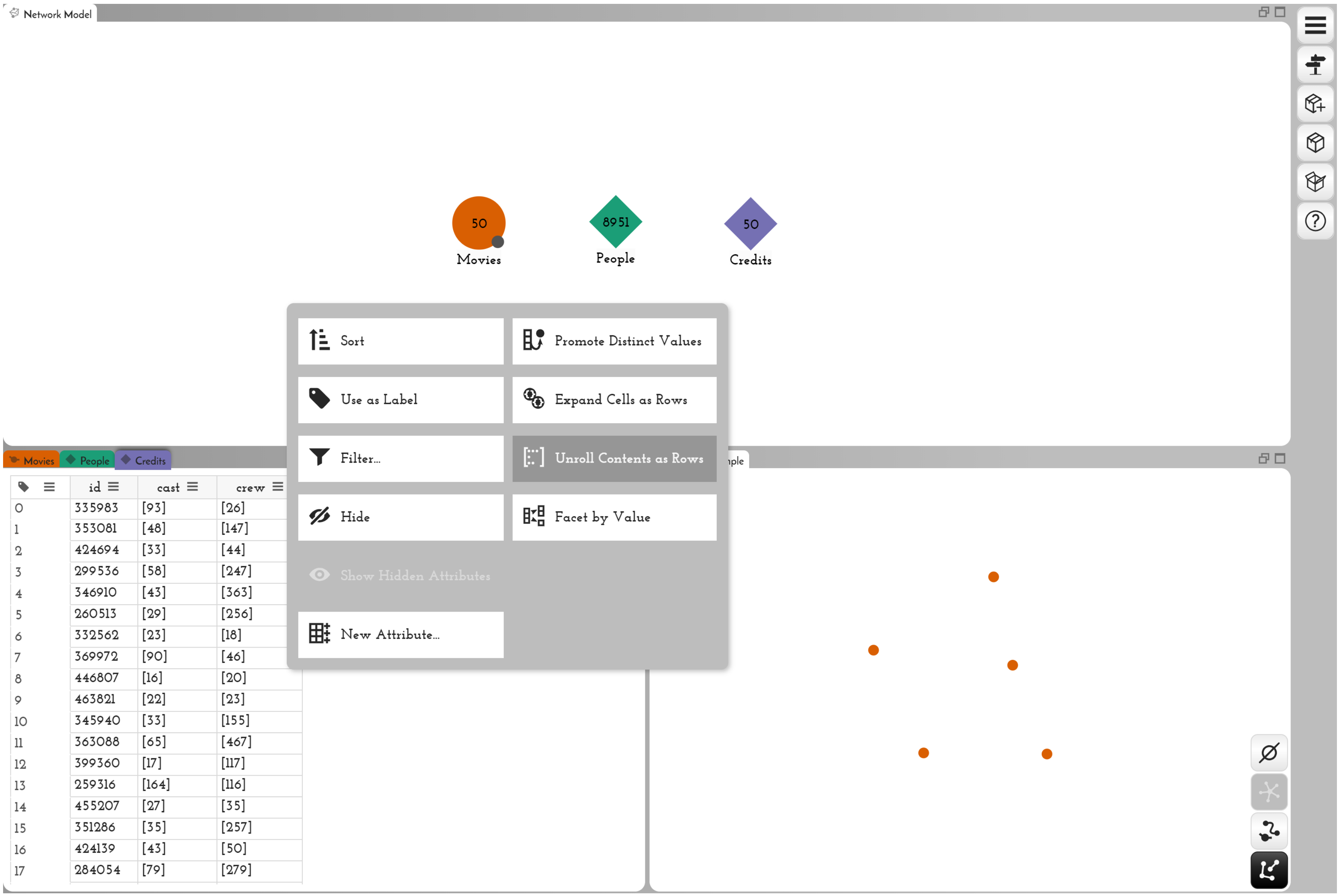

Figure M7: Here, we unroll nested crew structures (displayed as bracketed counts in the table) as a distinct crew class.

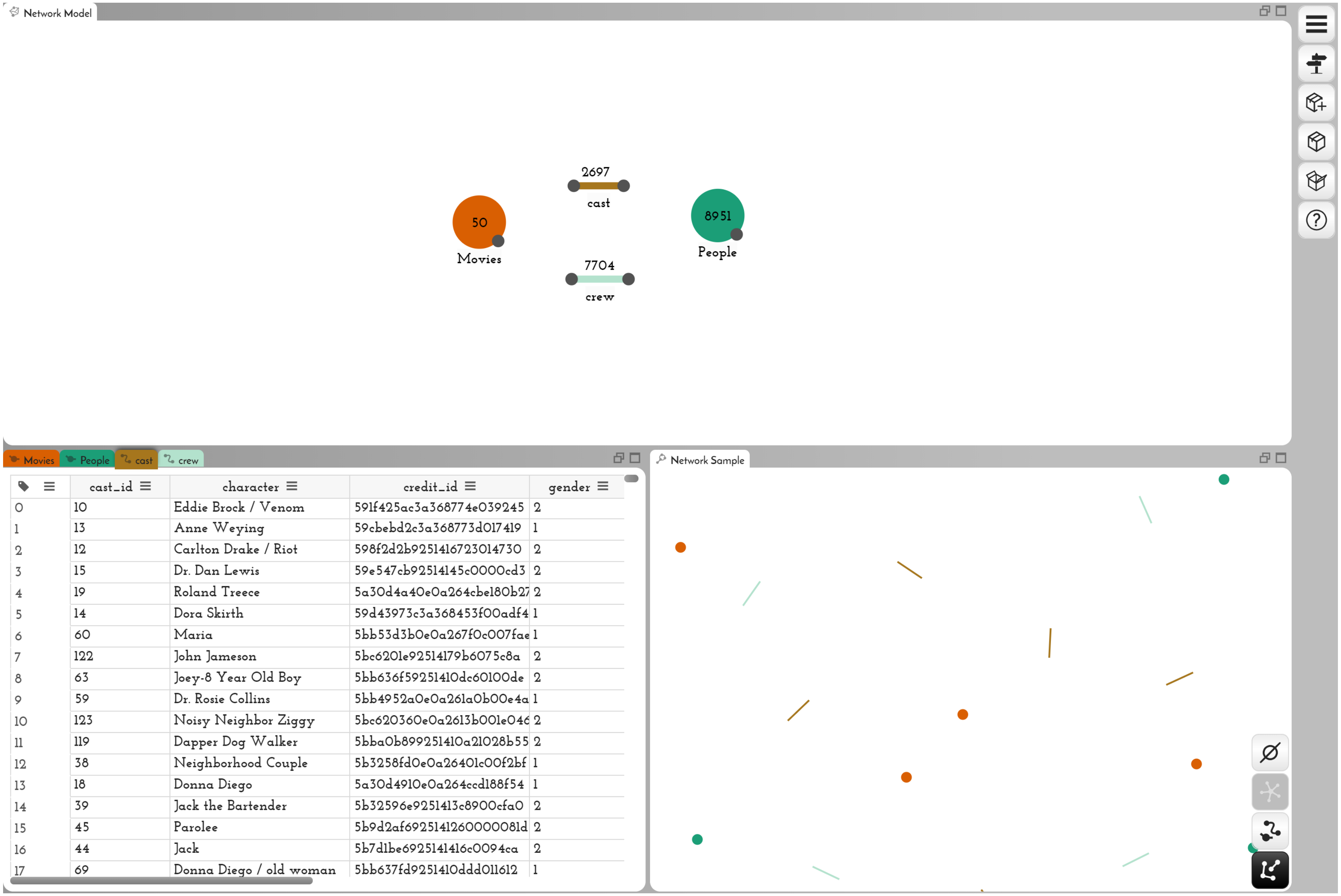

Figure M8: The state of the network model after crew and cast have been unrolled.

Figure M9: The connection interface, where users select a pair of attributes to join, in order to establish a connection between classes. Histograms at the top and bottom indicate how many connections per item will be created. The stacked bar chart in the center shows how each class contributes to an overall heuristic (see Section 6.4 in the paper) that is suggestive of which pairs of attributes may be a sensible match. The tables on the right are linked to provide context, giving a preview of the values that will be matched with each other.

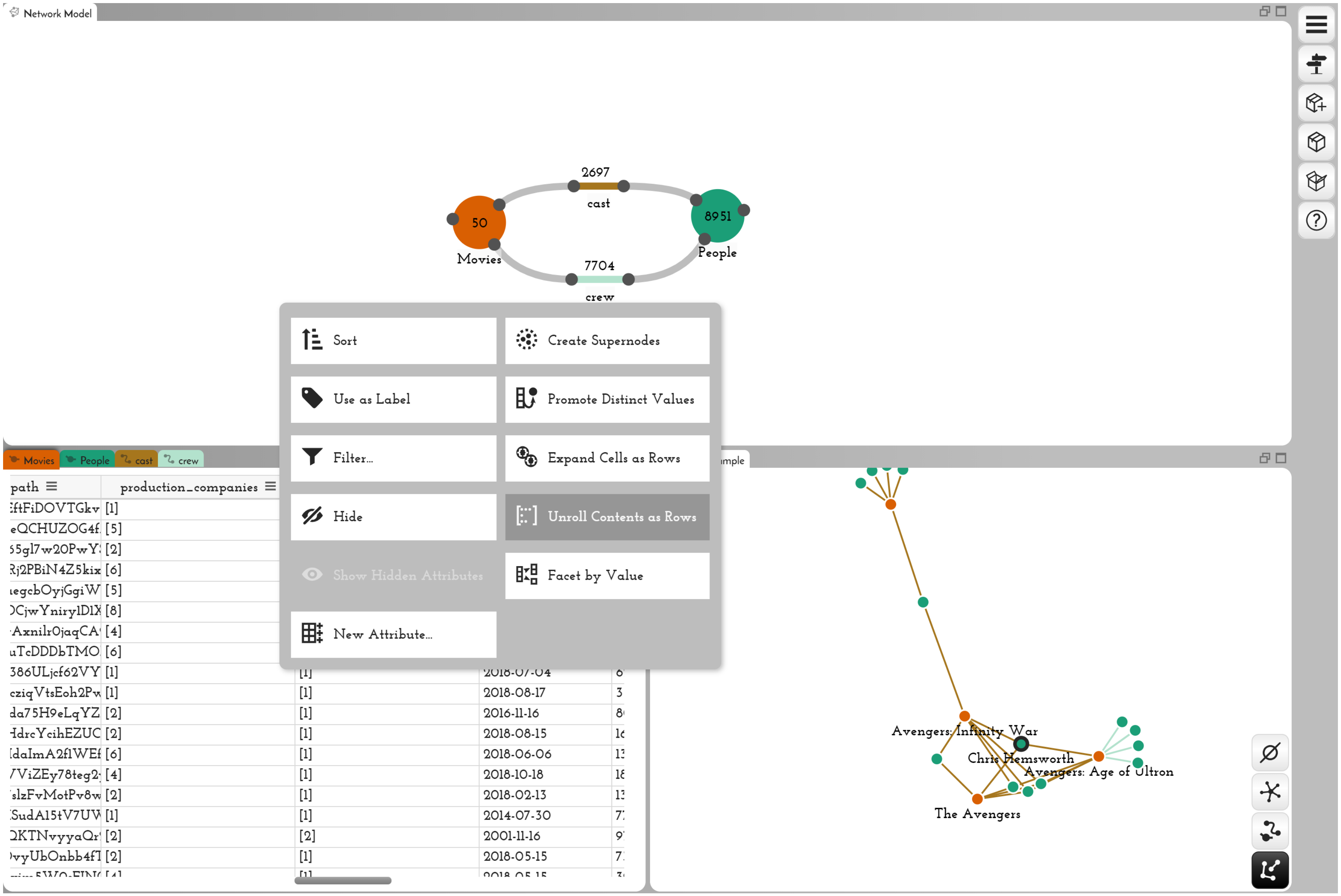

Figure M10: Here we unroll nested production company structures; as before, the count of nested objects within each list is displayed in brackets in the table.

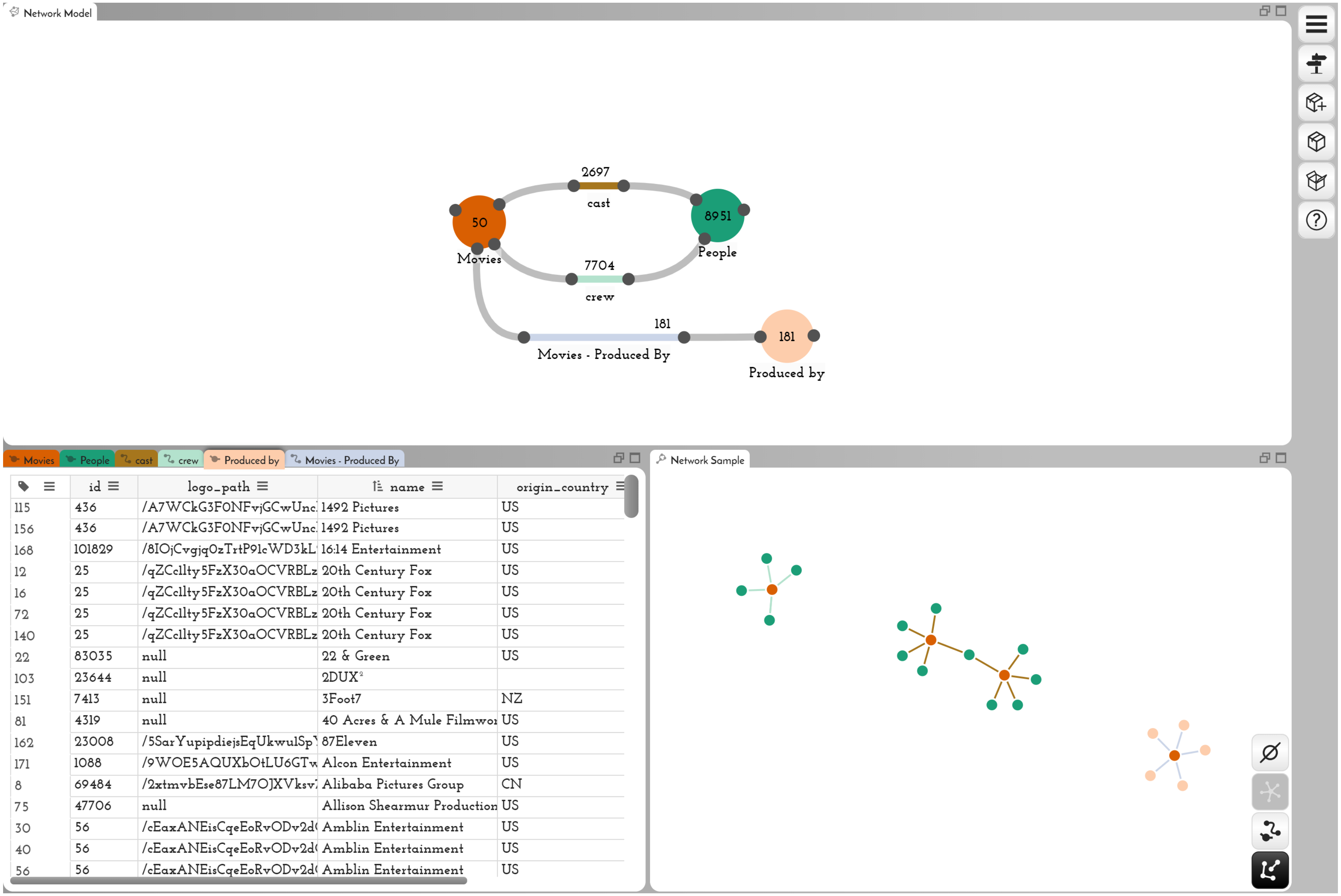

Figure M11: With production company objects unrolled, we can see that there are many duplicates in the name column—these are redundant items from the original dataset that need to be combined.

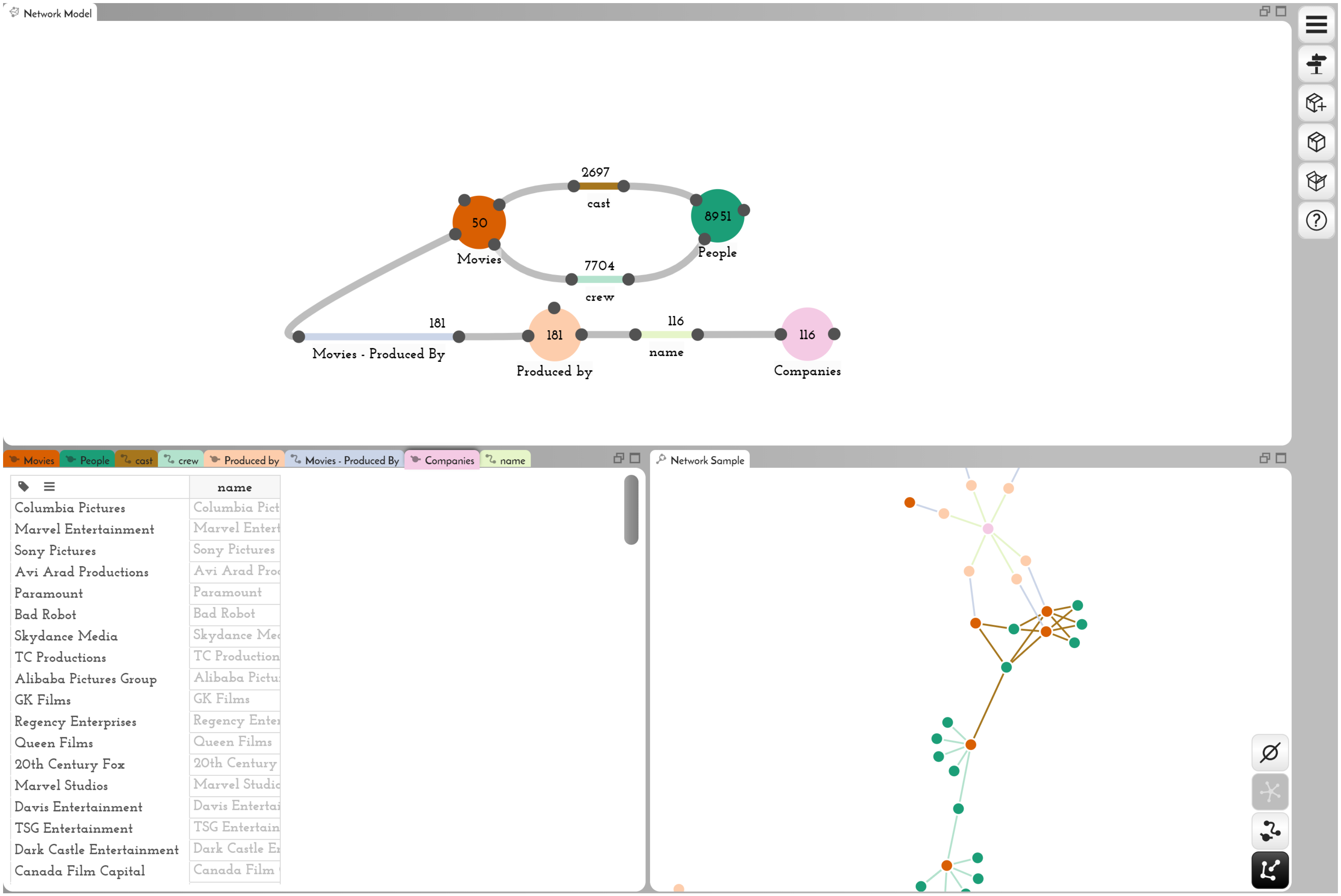

Figure M12: The state of the network after promoting the name attribute of the duplicate items.

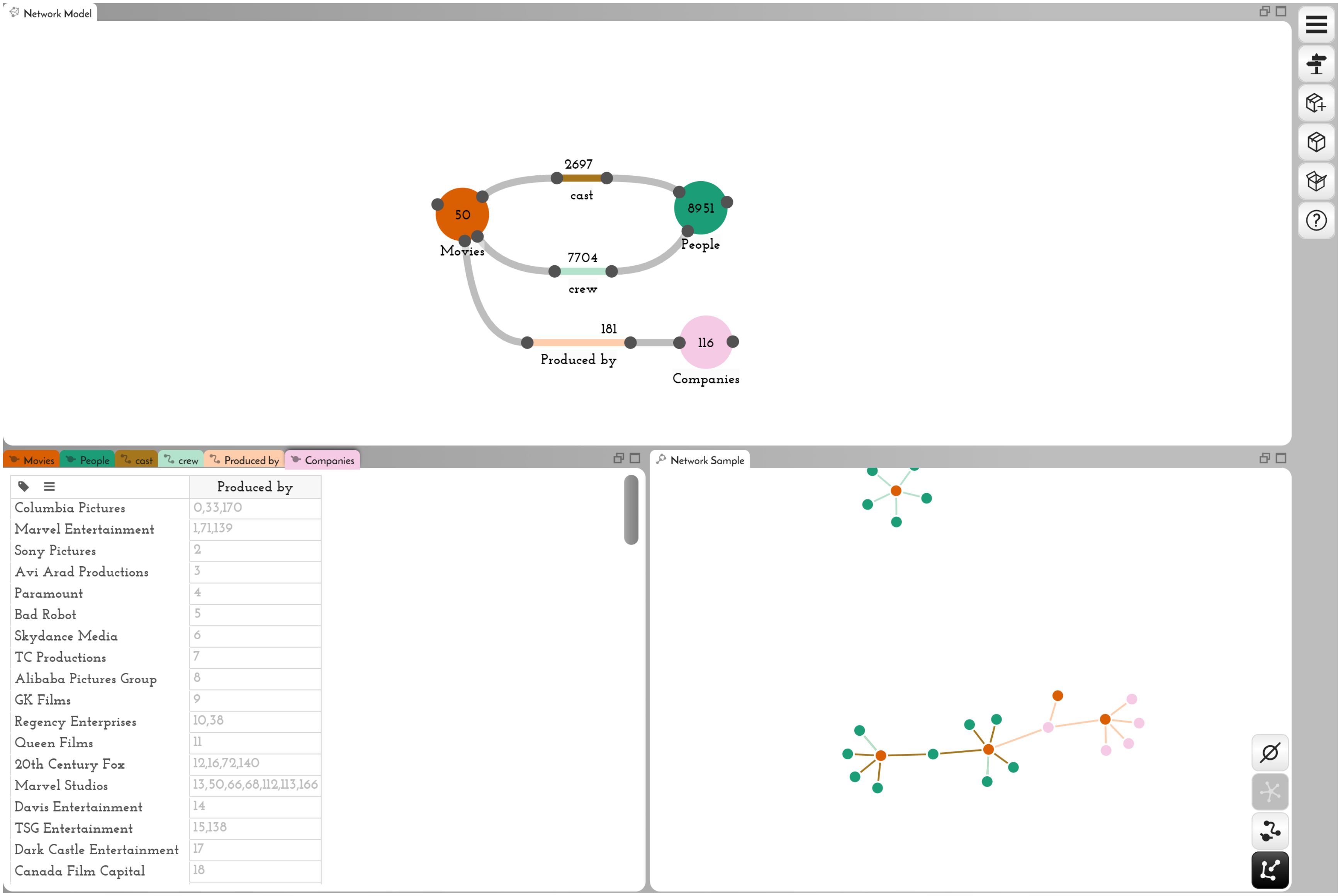

Figure M13: Here, we have converted the intermediate, redundant items into edges, for a cleaner network model.

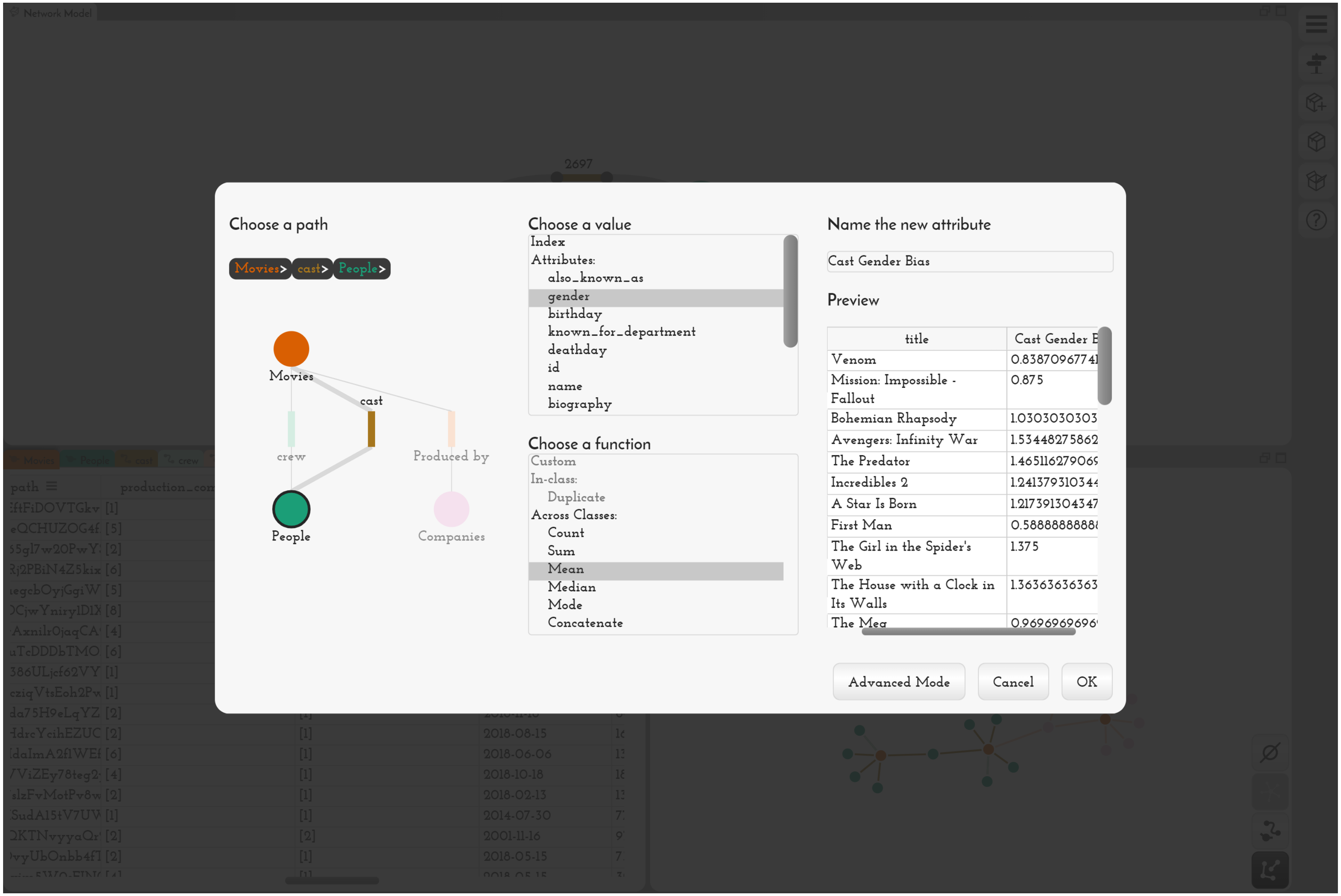

Figure M14: We wish to better understand gender bias in Movies, however, the gender attribute only exists on People. Here we begin to derive a Cast Gender Bias metric through cast edges, first by choosing the mean calculation as a template that we will adapt.

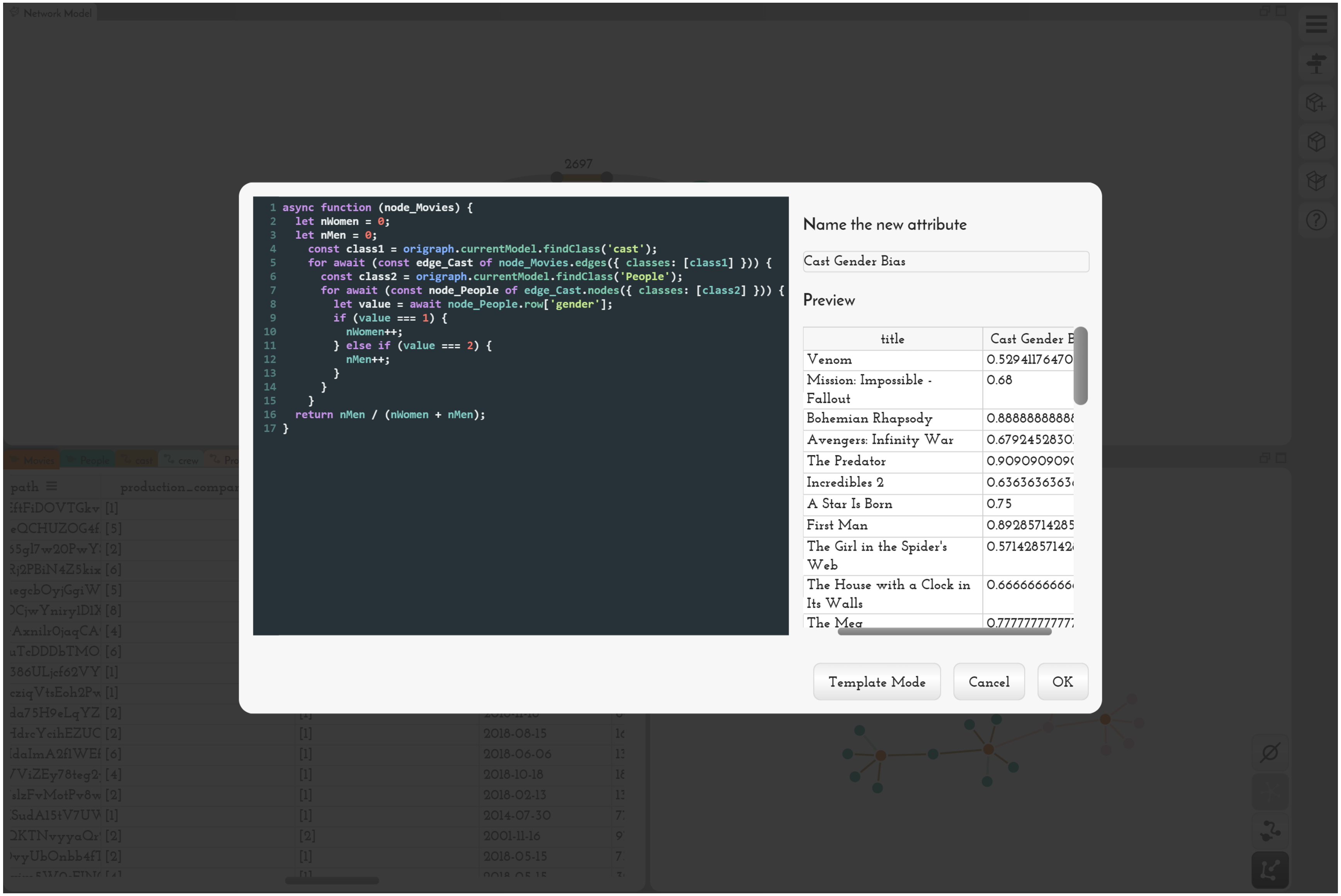

Figure M15: To calculate gender bias, we adapt the auto-generated code for mean, changing eight lines of code (2-3; 9-13; 16) to derive a custom value.

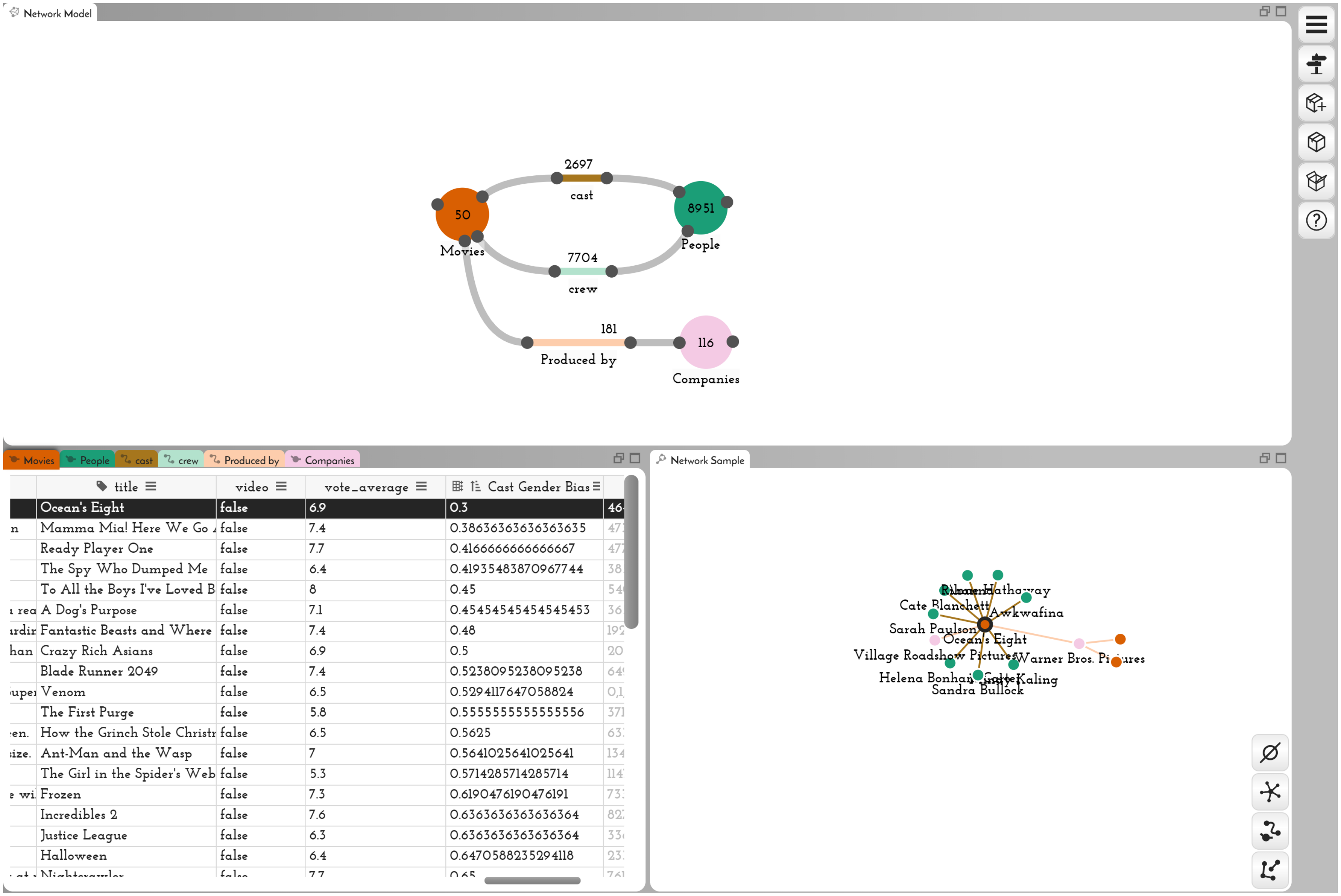

Figure M16: If we sort by our custom Cast Gender Bias, we see that Ocean's Eight has the lowest gender bias of the Movies in this dataset.

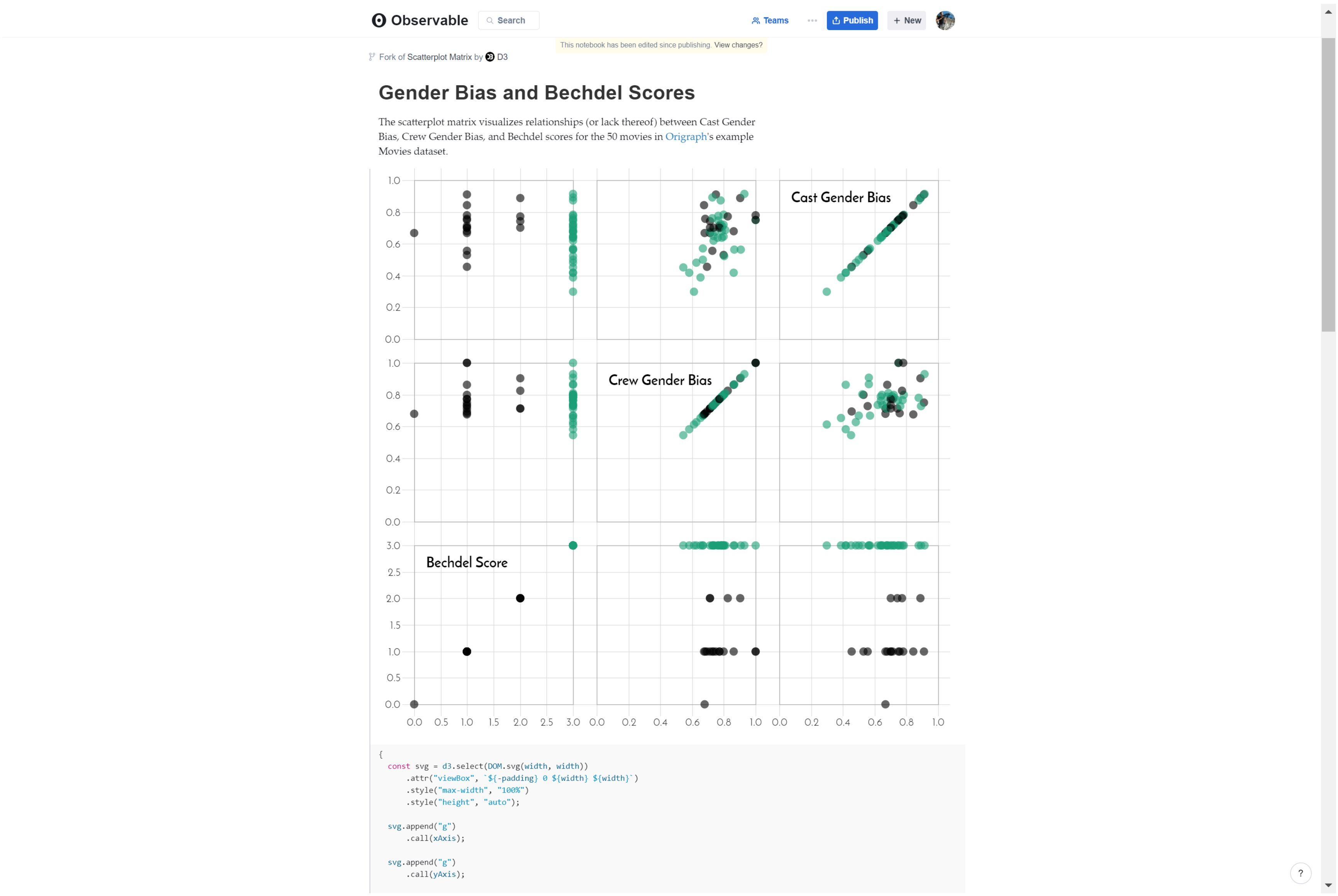

Figure M17: To better understand the relationships between Bechdel Scores, Cast Gender Bias, and Crew Gender Bias, we can export the Movies class as a CSV file, that we can visualize in an Observable Notebook scatterplot matrix. Green dots indicate movies that pass the Bechdel test.

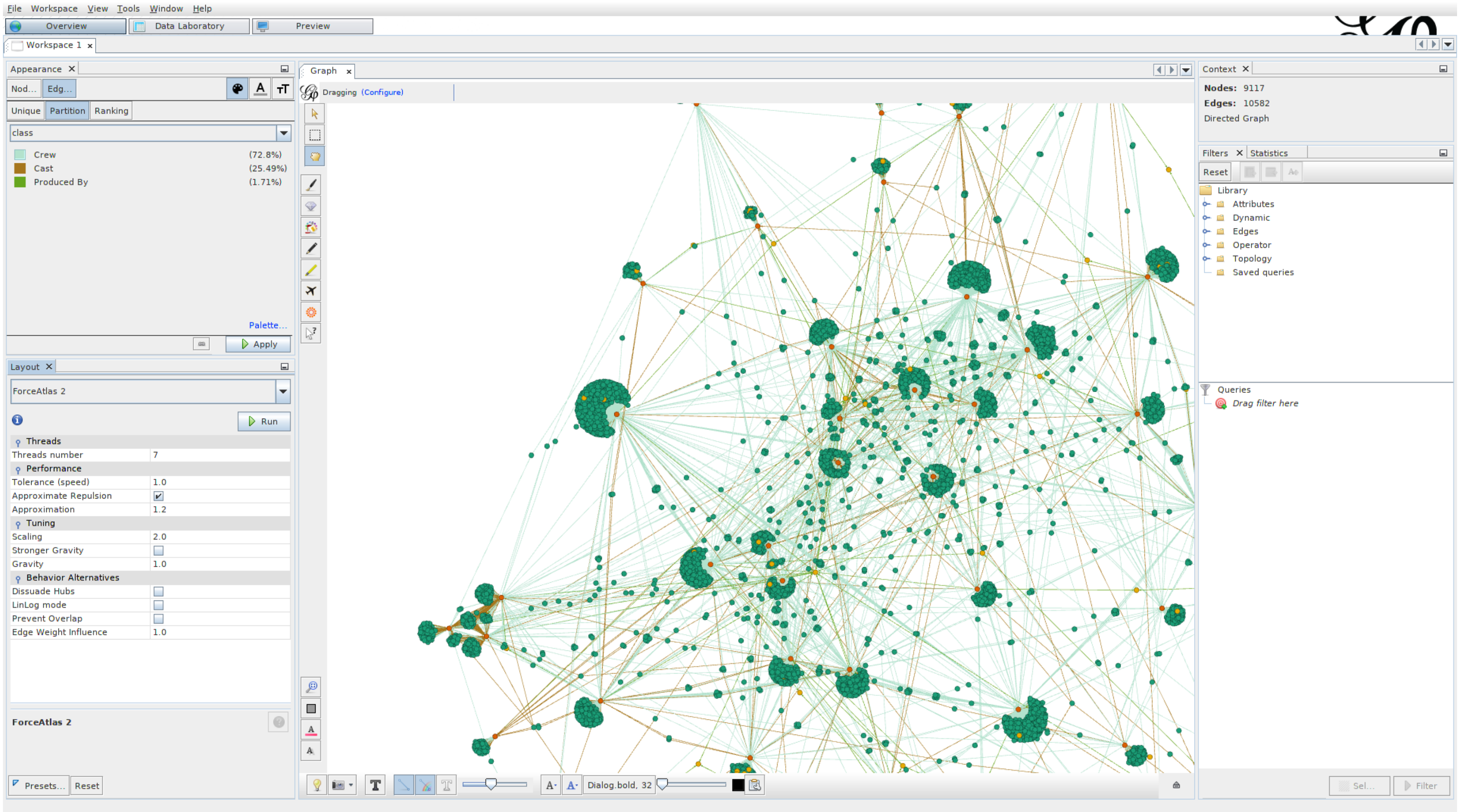

Figure M18: We can also export the network that we have modeled to other general-purpose visualization systems; in this case, we loaded the exported graph in Gephi.

3 Money and Political Support for the War in Yemen

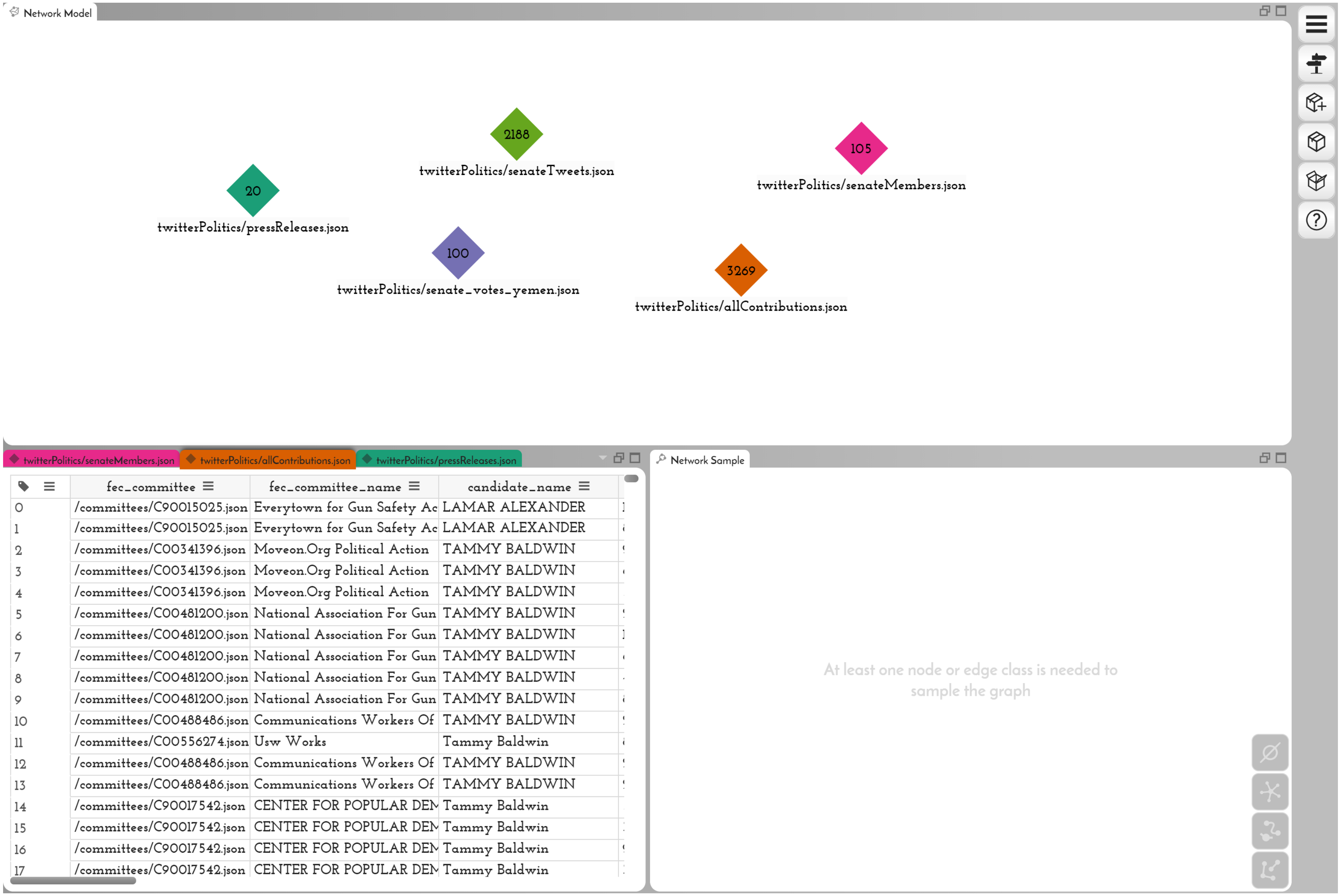

Figure Y1: In this use case, described in Section 9.2 in the paper, we combine data from ProPublica [5] and Twitter [7] to investigate the relationship of donors and US senators, specifically how senators voted about the war in Yemen.

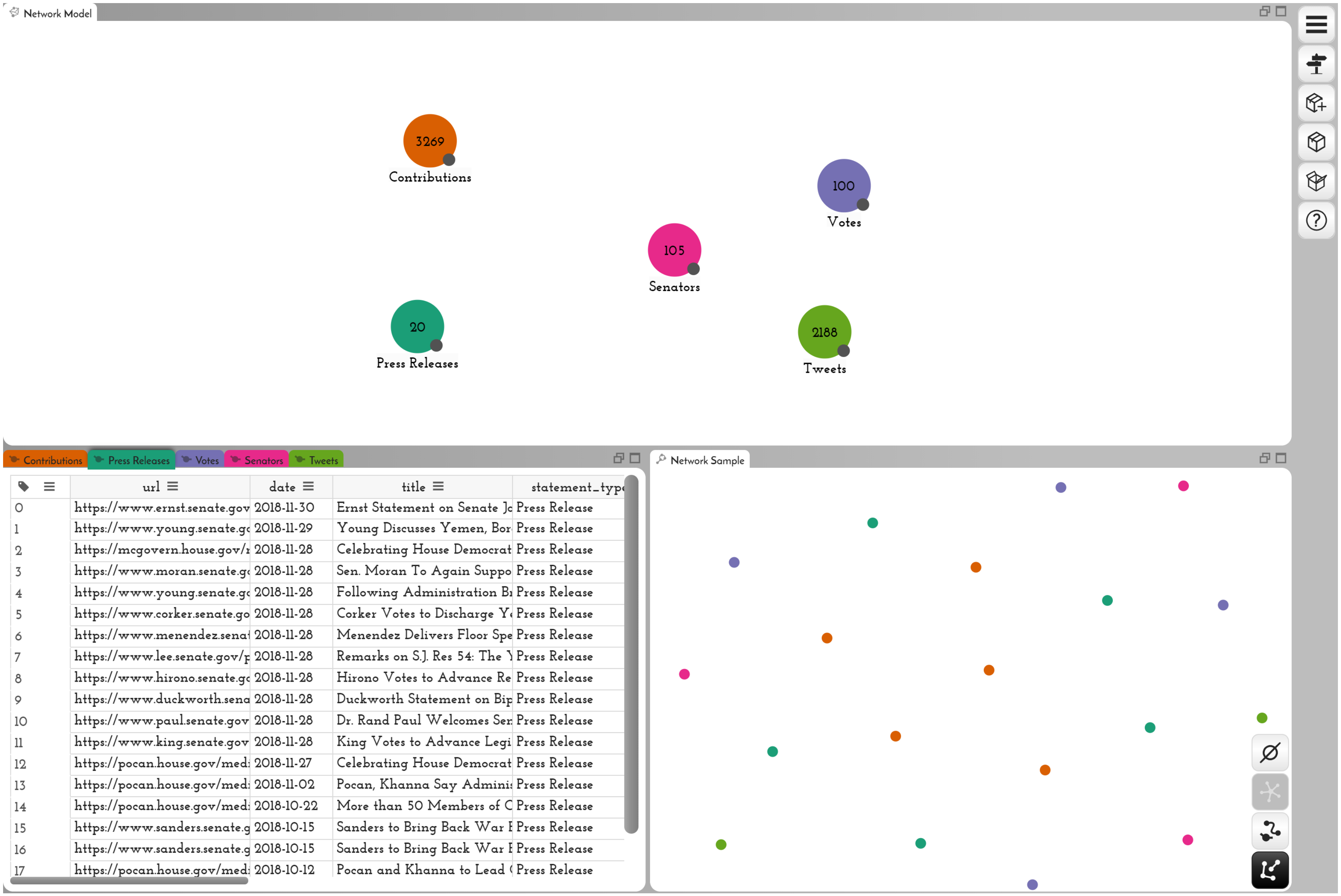

Figure Y2: We begin by converting each raw table to a node class.

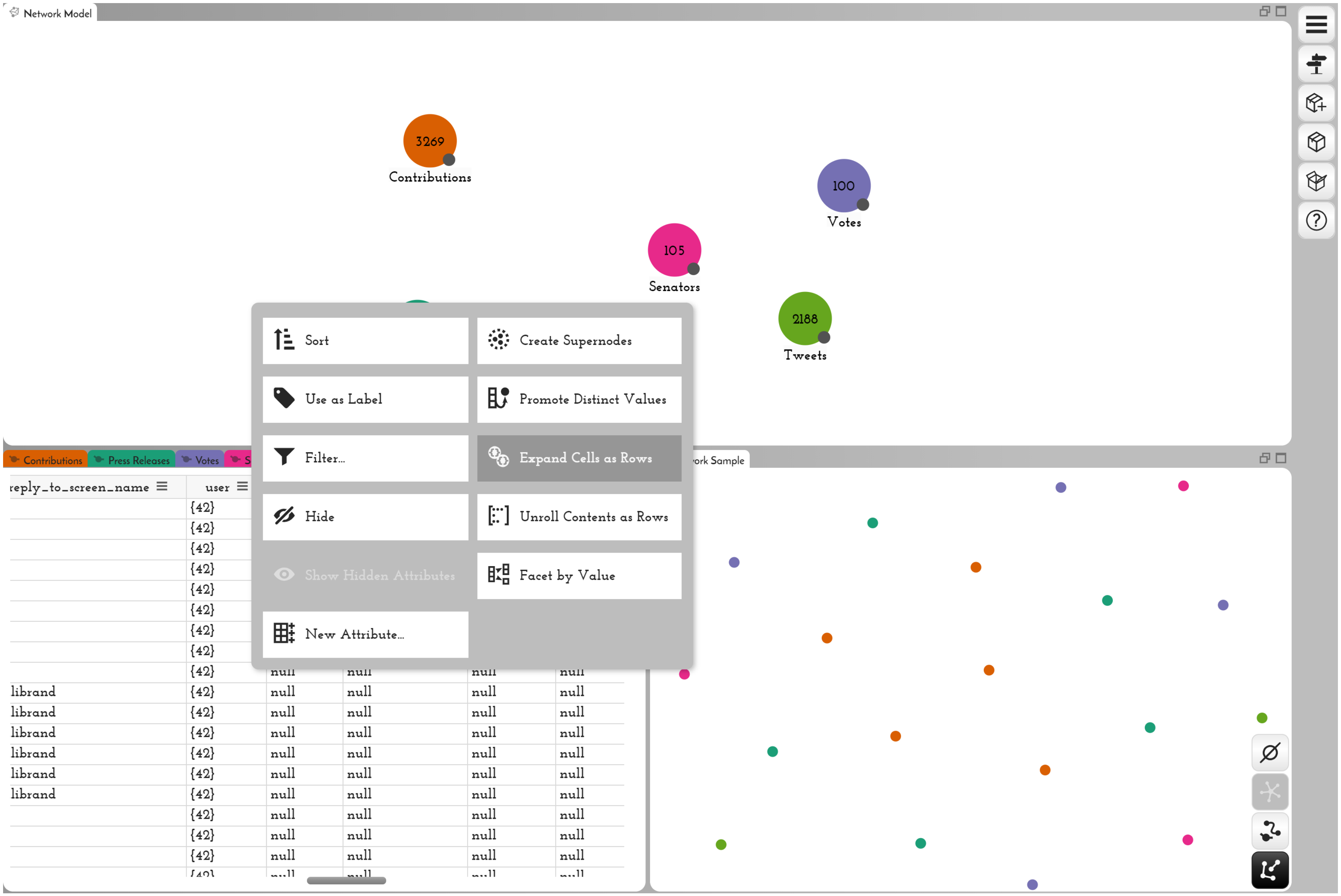

Figure Y3: To connect individual tweets to senators, we first expand nested user information. The curly braces in the cells indicate that a single object is nested in each row, with 42 attributes of its own.

Figure Y4: With user objects expanded, we can see there are many duplicates—there is a distinct user object for each tweet. At this stage, we could promote repeated values, however our objective is only to connect tweets to senators.

Figure Y5: To connect senators to tweets, we match their `screen_name` attribute from the Twitter user nodes to the `twitter_account` attribute of ProPublica Senator nodes.

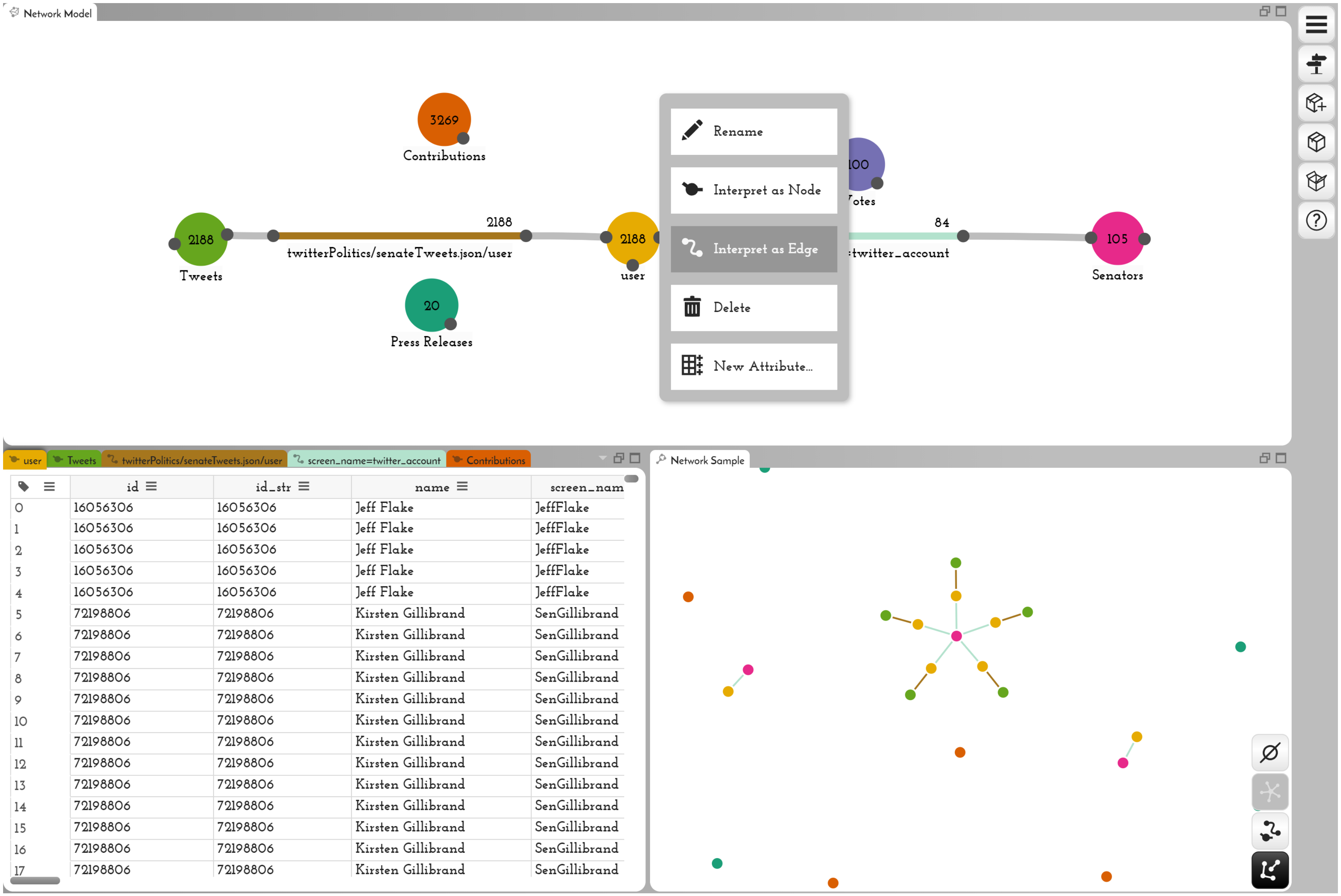

Figure Y6: We clean up the network by converting the intermediate user nodes to edges.

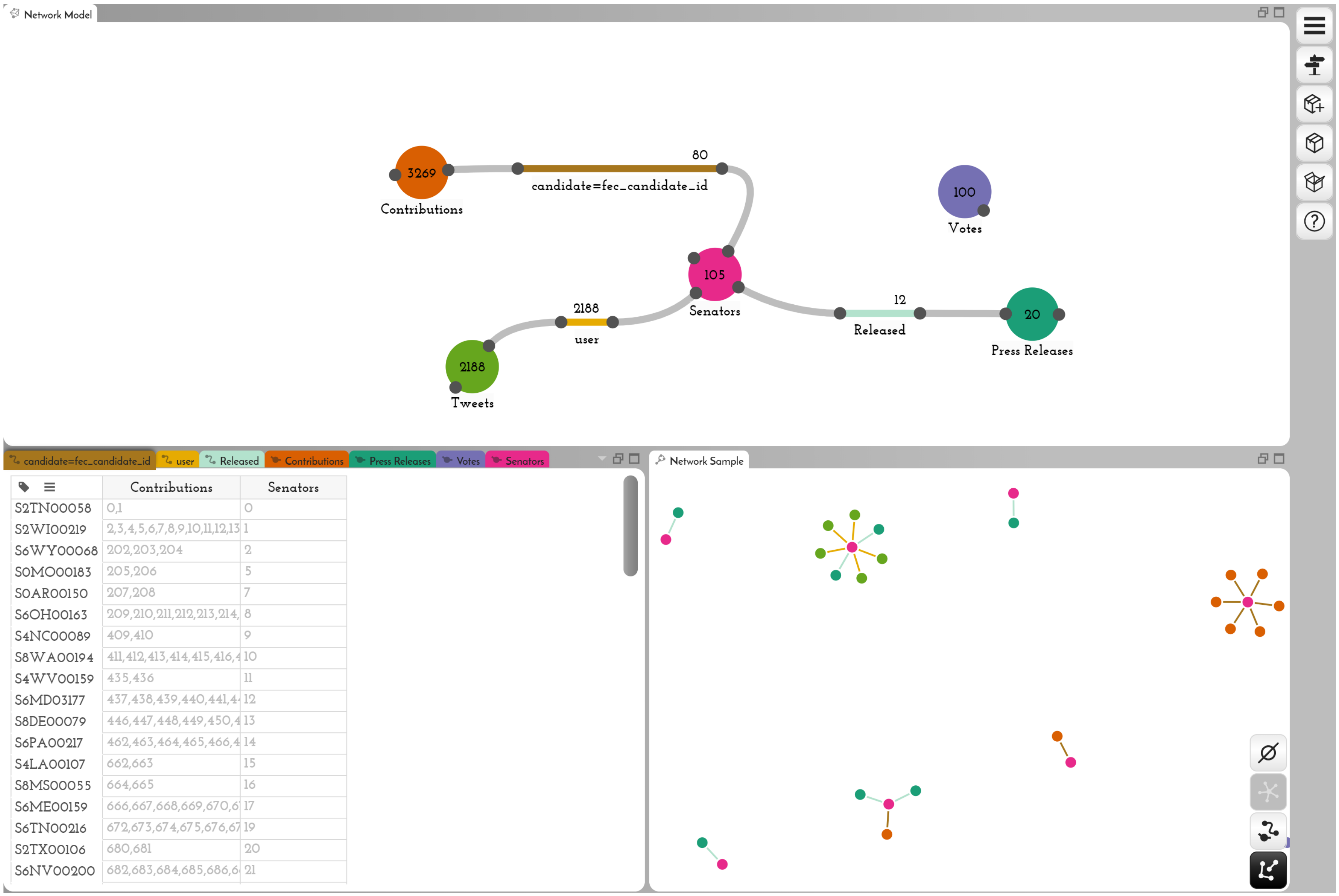

Figure Y7: With tweets connected, we add also connect the Contributions and Press Releases classes based on their native IDs.

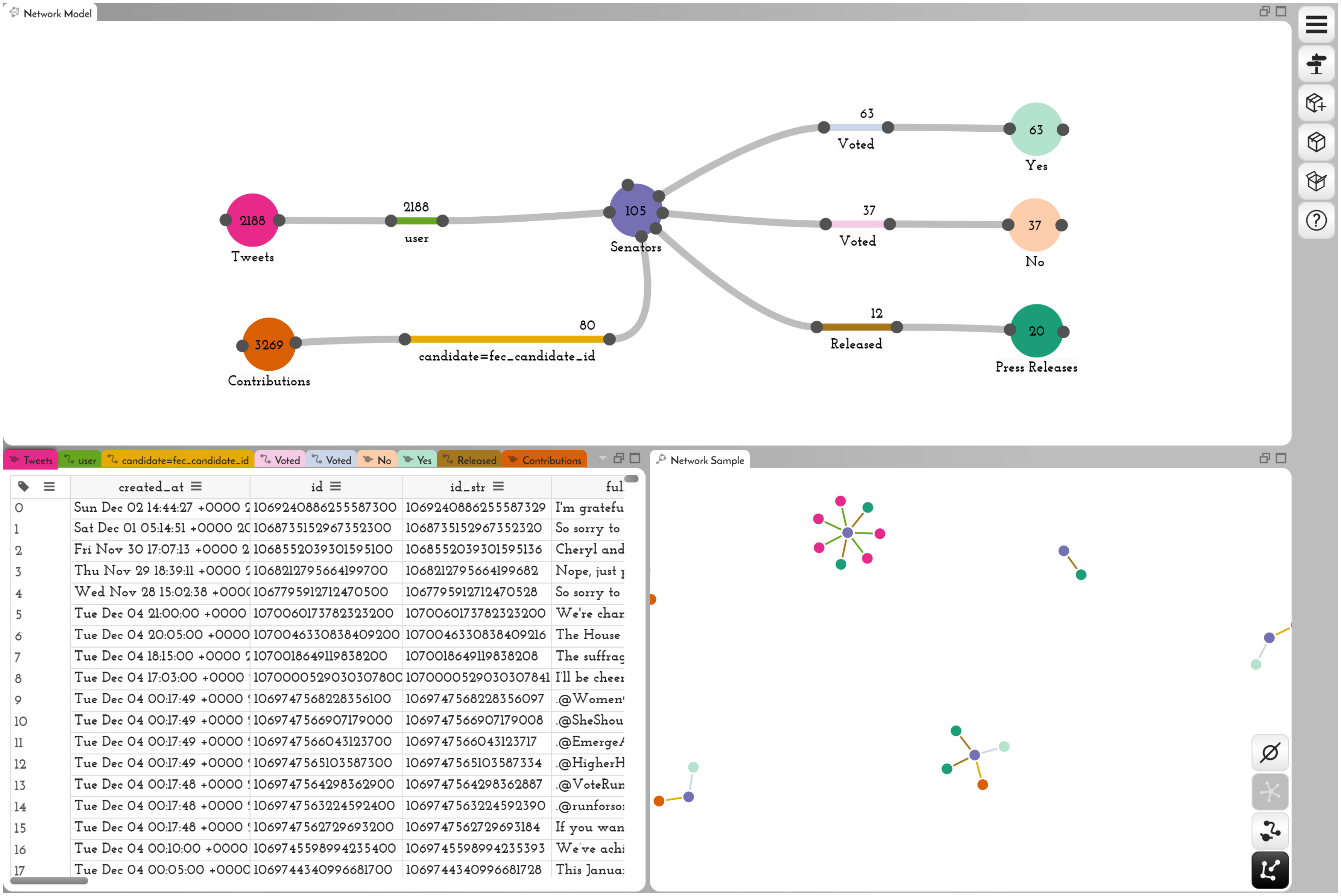

Figure Y8: As we are interested in how Senators voted, we facet the Votes class before connecting them to Senators.

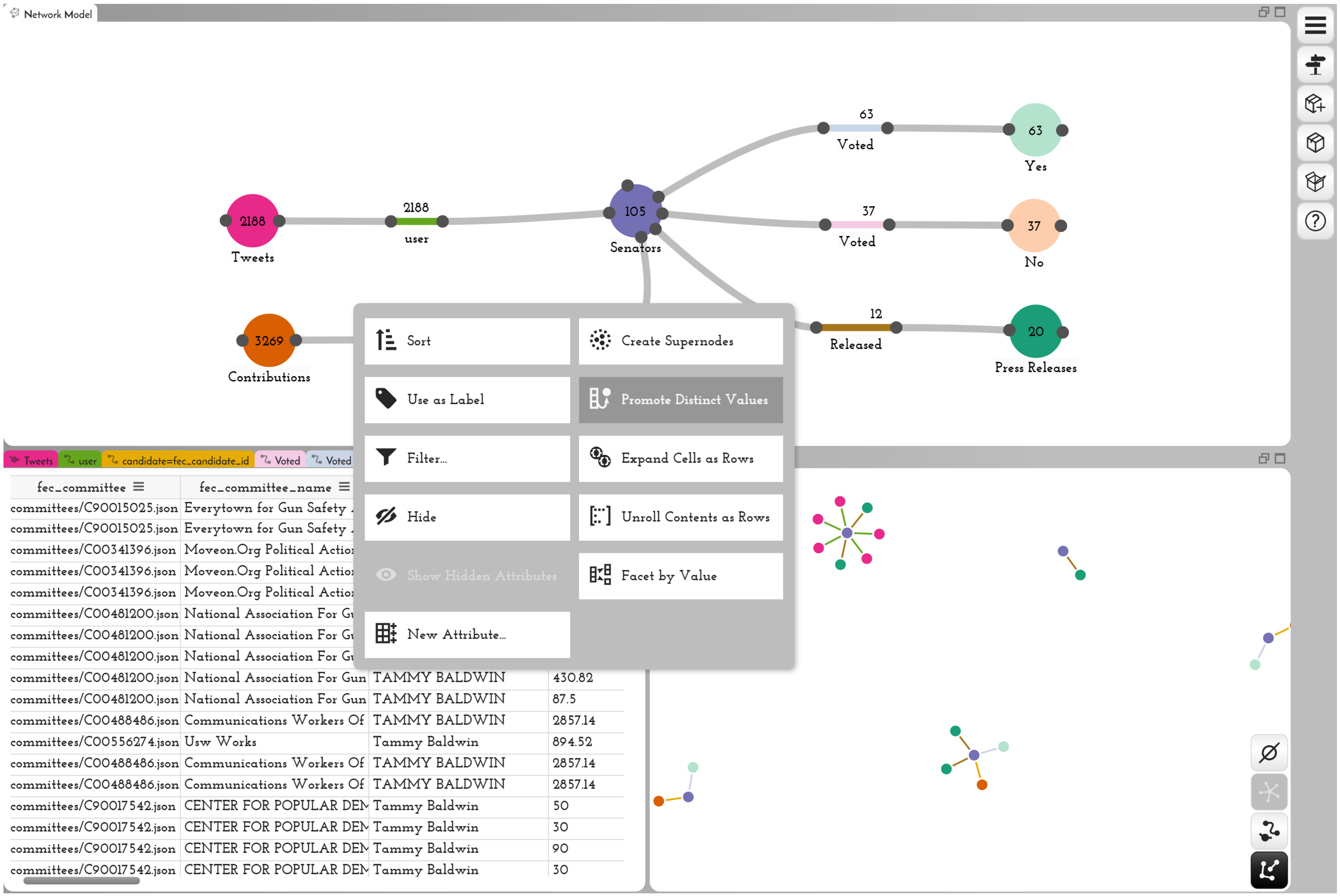

Figure Y9: We are also interested in donors, and their relationships to Senators—however, the Contributions table contains many rows from the same donor. To identify individual donors, we promote the `fec_committee_name` attribute.

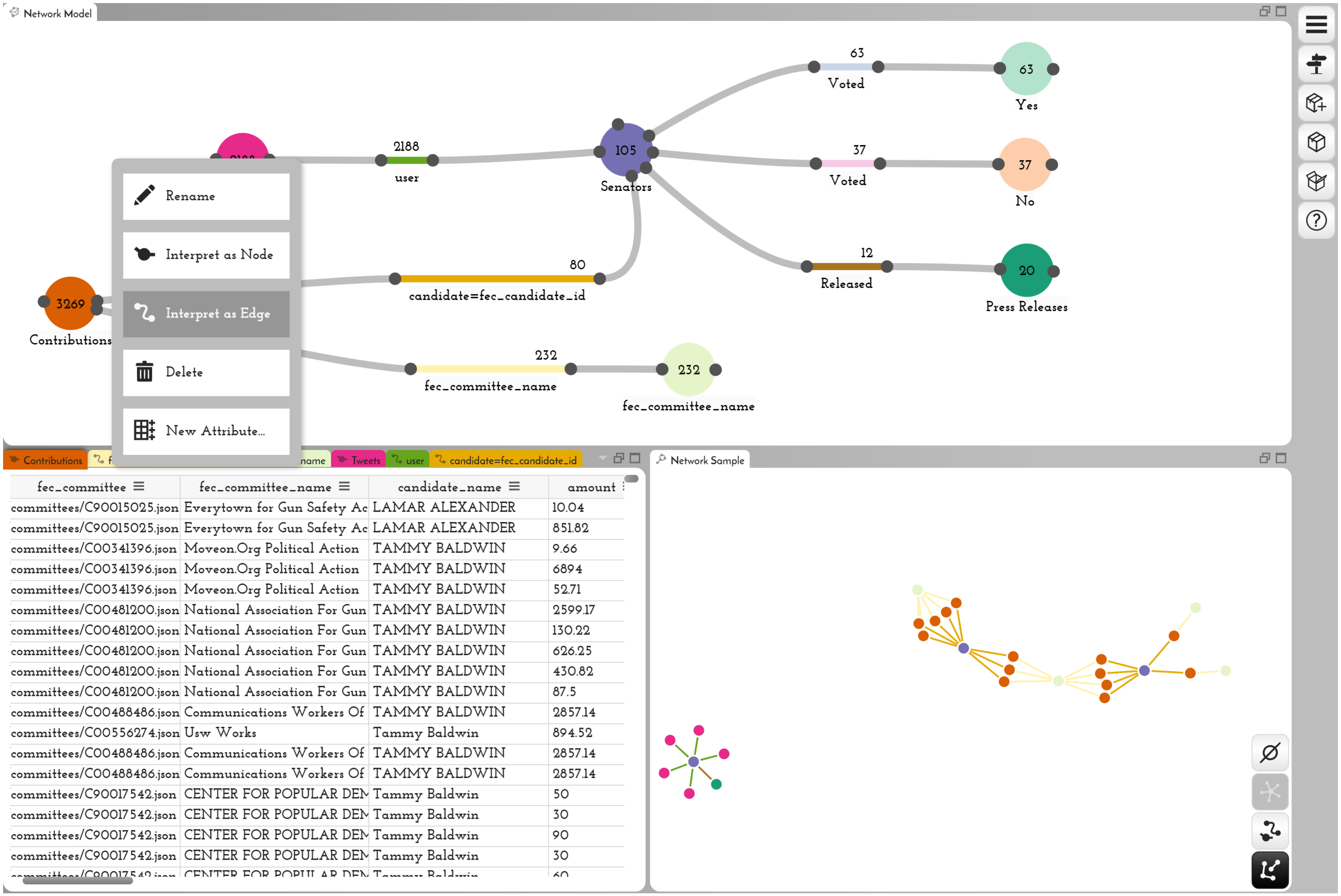

Figure Y10: To clean up the network model, we now convert the Contributions class to an edge class.

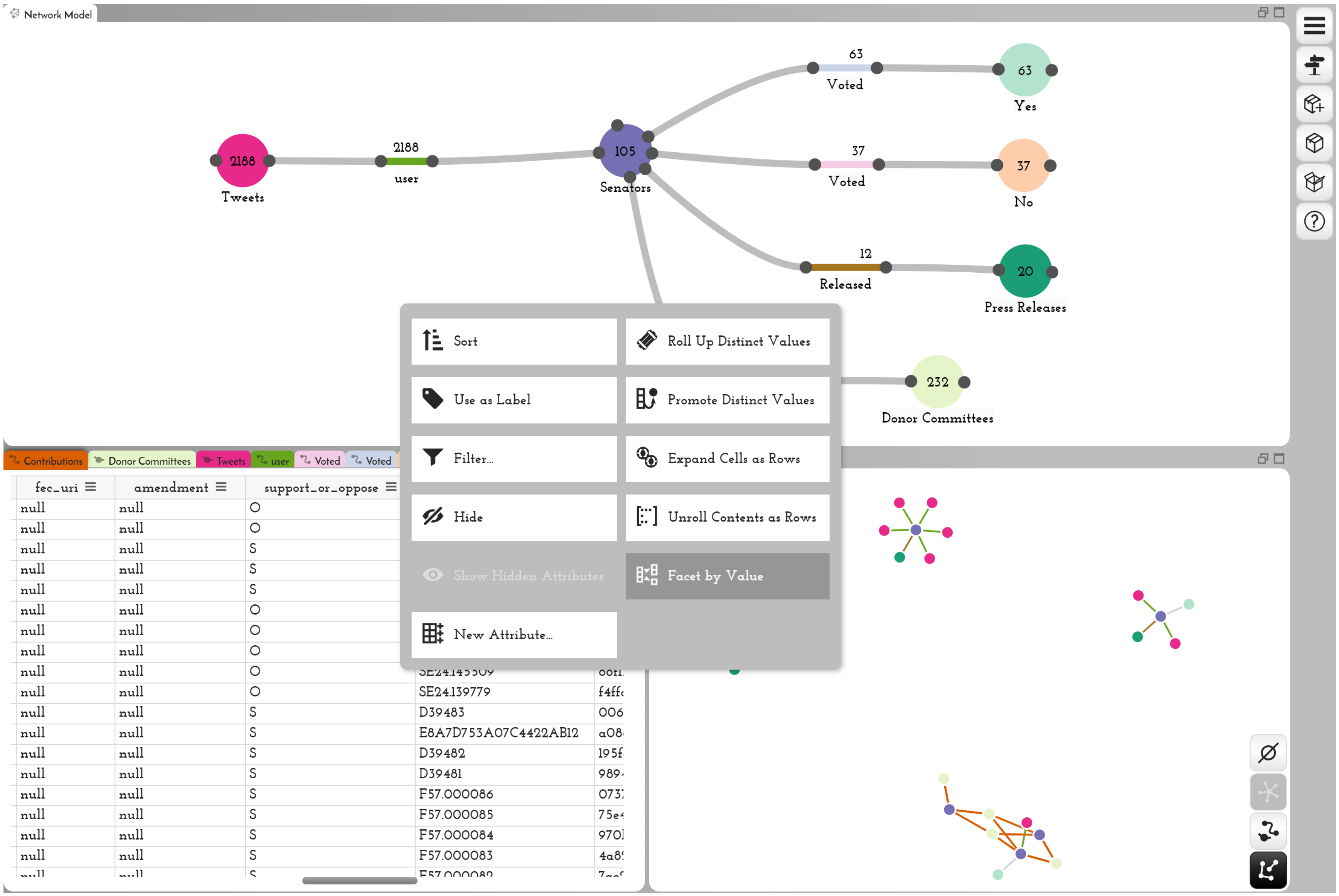

Figure Y11: Political contributions can be made both in favor, or in opposition to, a Senator. To distinguish between these different relationships, we facet the Contributions edge class.

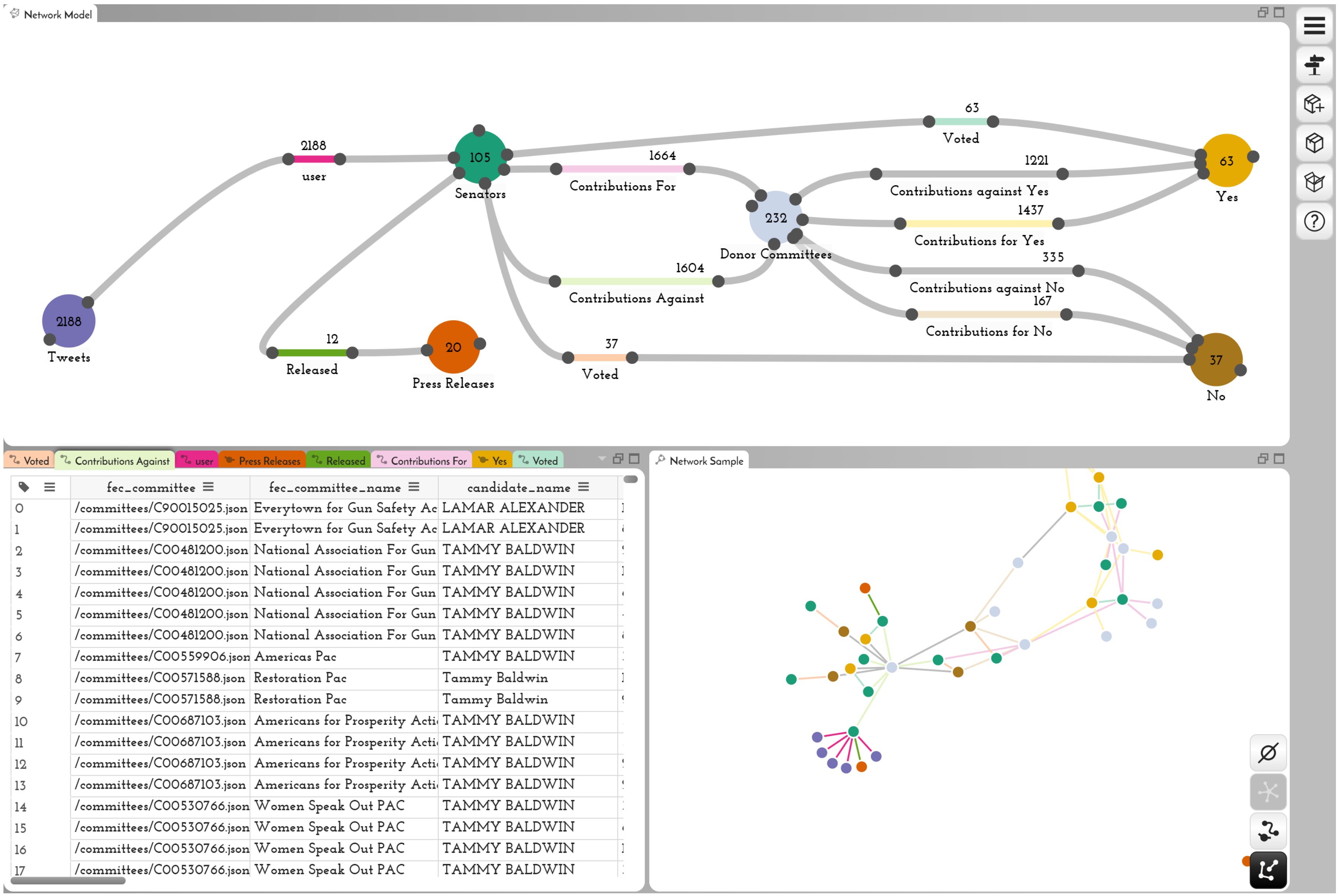

Figure Y12: If we wish to model relationships between donors and the votes on the Yemen bill, we can project edges that bypass Senators, connecting donors and votes directly. Here, we produce four new edge classes: links between donors that financially supported, or opposed, votes in favor, or against the bill.

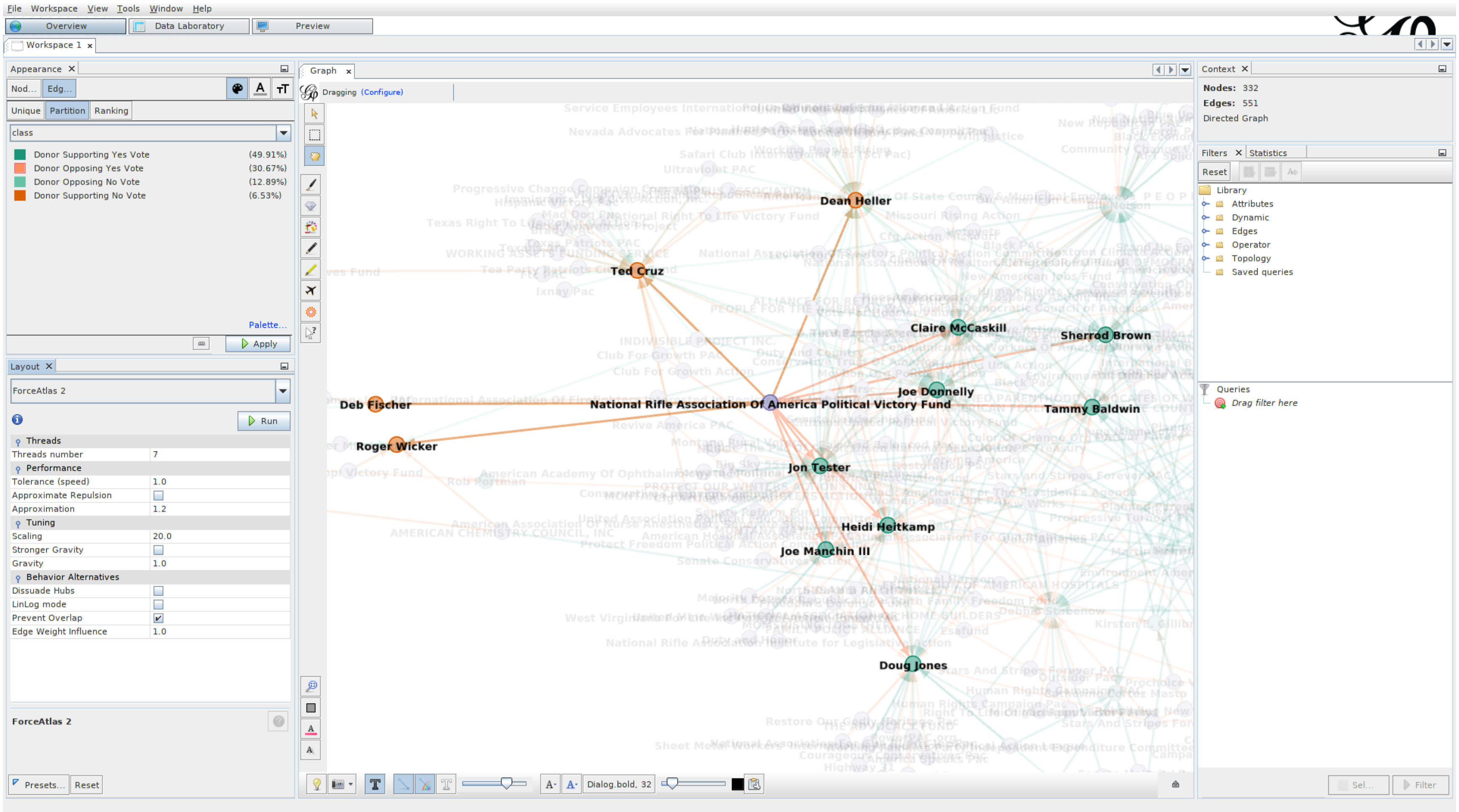

Figure Y13: Projected edges, donors, and votes are exported to Gephi. We highlight the NRA, as it did not donate to any senators who voted Yes—only to those who voted No, and donated in opposition to other senators who voted Yes.

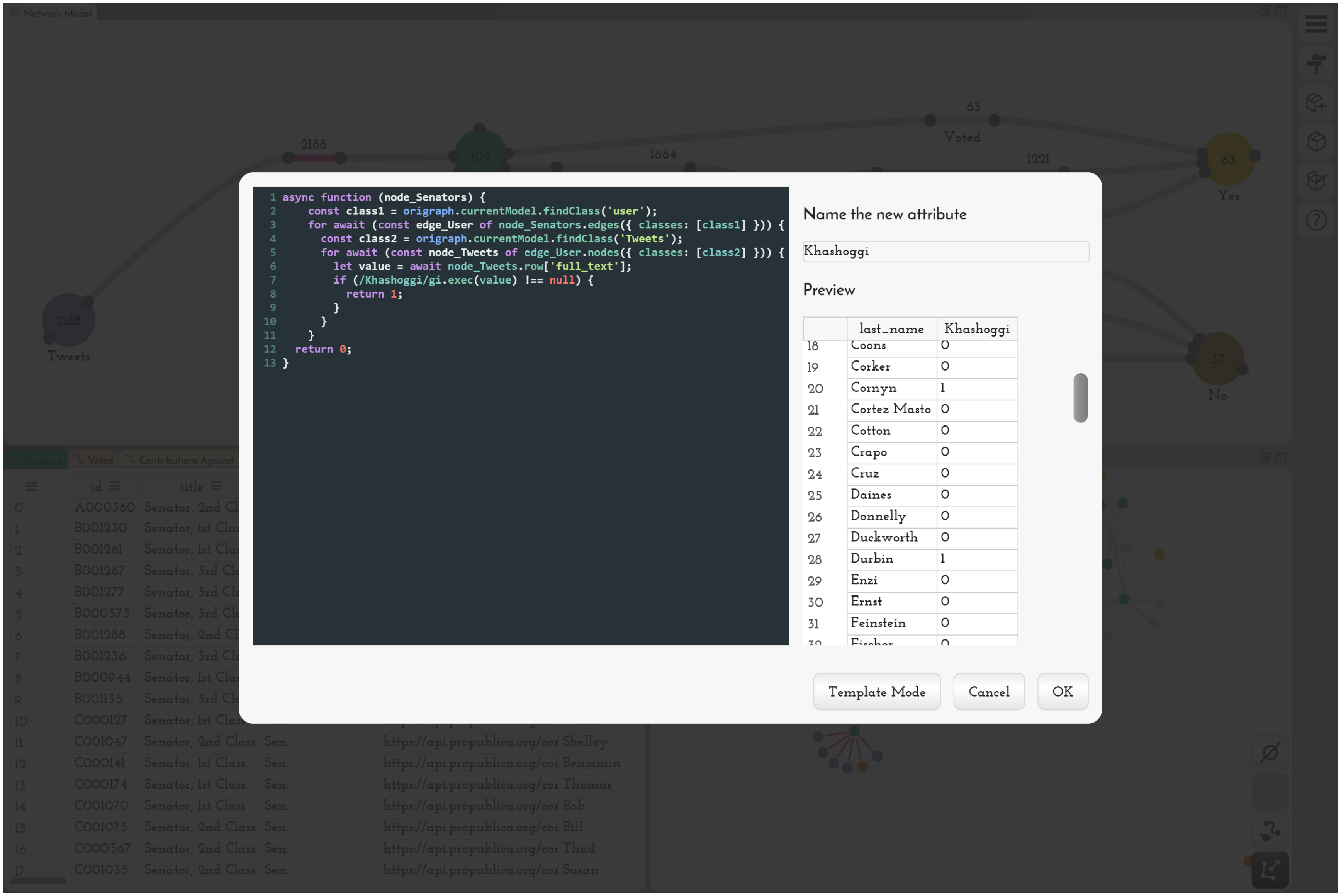

Figure Y14: To ascertain which Senators were more or less vocal about the issue, we can derive attributes that indicate whether or not strings of interest appeared in their tweets or press releases. In this instance, we compute whether the string "Khashoggi" appears in the text of a Senator's tweet.

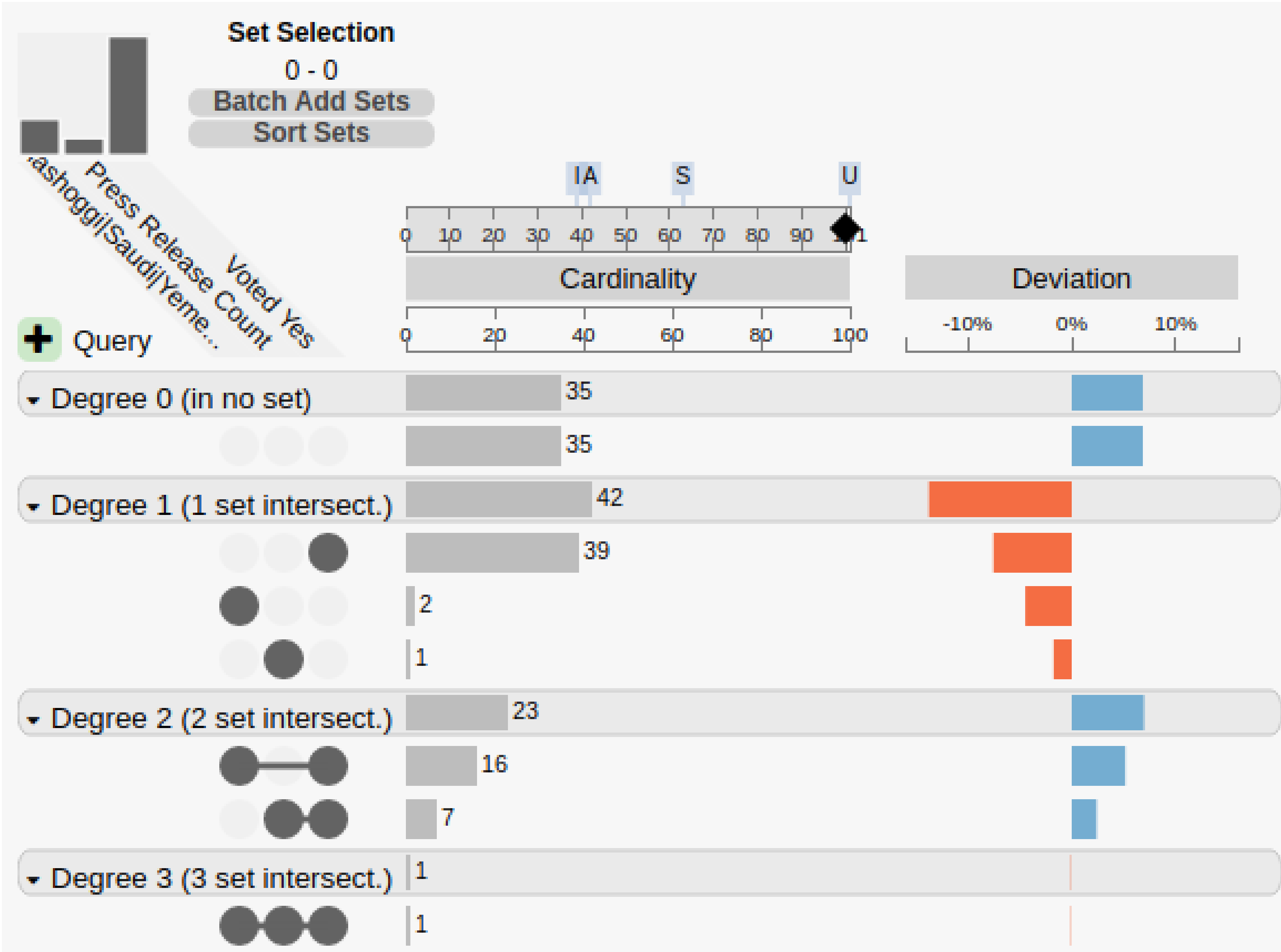

Figure Y15: Whether senators tweeted about, issued press releases, and/or voted Yes, visualized in UpSet [3]. We can see that, with the exception of only three senators, all senators who voted No were relatively silent about their votes.

References

[1] T. M. D. Community. The Movie Database (TMDb). https://www.themoviedb.org/documentation/api, 2018.

[2] J. Heer and A. Perer. Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. *Information Visualization*, 13(2):111–133, Apr. 2014.

[3] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, Dec. 2014.

[4] Z. Liu, S. B. Navathe, and J. T. Stasko. Ploceus: Modeling, visualizing, and analyzing tabular data as networks. *Information Visualization*, 13(1):59–89, Jan. 2014.

[5] ProPublica. The ProPublica Congress API. https://www.propublica.org/datastore/api/propublica-congress-api, 2018.

[6] B. Test. The Bechdel Test Movie List. https://bechdeltest.com/, 2018.

[7] I. Twitter. Twitter Developer Platform. https://developer.twitter.com/content/developer-twitter/en.html, 2018.