Episogram: Visual Summarization of Egocentric Social Interactions

Nan Cao ▪ *TongJi University*

Yu-Ru Lin ▪ *University of Pittsburgh*

Fan Du ▪ *University of Maryland*

Dashun Wang ▪ *Northwestern University*

The interactive visualization tool Episogram summarizes the social interaction process based on a dynamic tripartite network and visualizes users' social behaviors by displaying and aggregating the network along multiple temporal dimensions, from different actors' egocentric perspectives.

Social interaction refers to a "dynamic, changing sequence of social actions between individuals (or groups) who modify their actions and reactions due to the actions by their interaction partner(s)" (http://en.wikipedia.org/wiki/Interaction). Datasets that archive such individual social interactions have become increasingly available. Examples include the content generated by hundreds of millions of users on social media such as Twitter, the communications and transactions recorded in emails and instant messages, and publications that document the collaboration among authors. These social traces provide abundant opportunities for understanding social interactions. For example, understanding the common features of users' communication activities helps analysts identify their common behaviors, thus helping to detect anomalous users, which is a serious need in the field of information security. However, understanding these data is not easy given the complexity of the datasets (which are often unstructured, dynamic, and heterogeneous) and the different types of social interactions in various application domains.

Data visualization enables the understanding complex data through intuitive representations, facilitating data interpretation and summarization. However, several challenges exist in visualizing the social interaction data. First, the activities that occur during social interactions (such as posting or retweeting) provide necessary context for understanding the meaning of the interactions.[1] Therefore, an efficient visualization should be able to display and capture such context-rich social interactions with a simple and integrative visual design. Second, designing a visualization that captures the temporal patterns (such as the frequency and duration of the social interaction process), content patterns (including the topics around which the interaction occurred), and behavior patterns (such as how a user posts or retweets in Twitter) is important for revealing insights. Furthermore, understanding about the common structure of social interaction processes is key overcoming these challenges.

In this article, we introduce a novel visualization design called *Episogram* for visualizing social interaction data based on an anatomy of the social interaction process in which the actors and objects involved can be formally represented as a time-varying tripartite network (see Figure 1). We begin here by providing an in-depth analysis of the key elements and structure of the social interaction process. Following this analysis, we introduce a directed tripartite network data model that can capture essential social interaction information in generalized social contexts. Our approach extends the Andrienko task model[2] to characterize differ-

 Published by the IEEE Computer Society

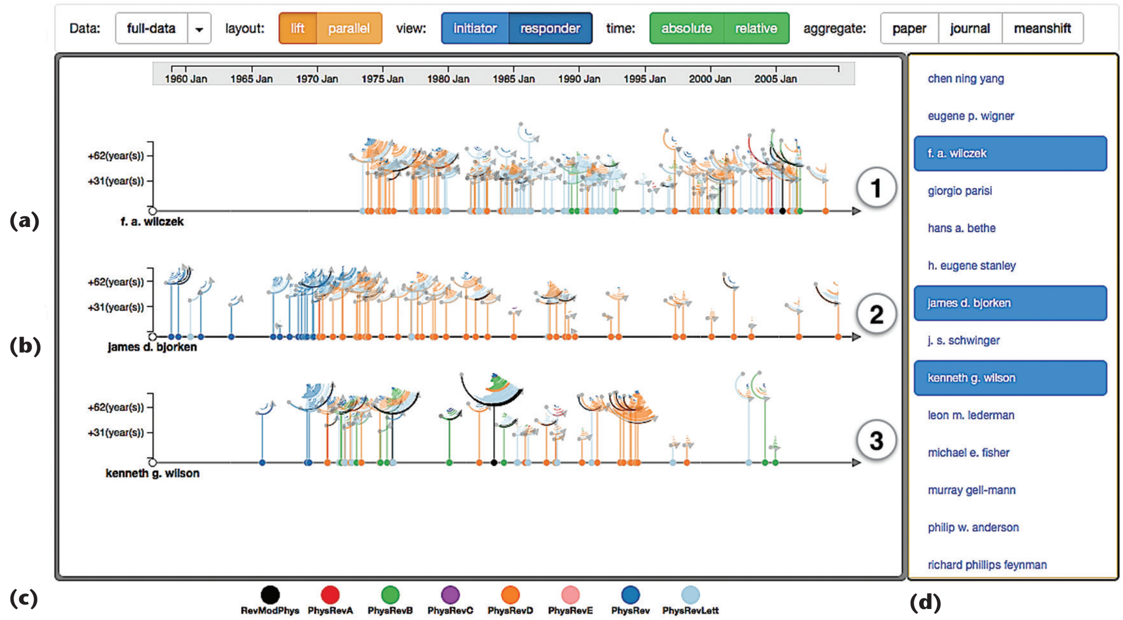

Figure 1. Episogram visualizations. This design helps users explore and compare social interaction data using egocentric viewpoints. In this case, the social interactions of the social actors (the three physics scholars) are visualized along the timelines. The social interaction events, including publishing papers (represented as vertical bars) and receiving citations (represented as crescent shapes on top of the vertical bars), are scattered according to when the events occurred. The major user interface components include (a) a toolbar, (b) the main display, (c) a legend, and (d) an actor list.

ent levels of user tasks in seeking information in social interaction data. Based on this task requirement, we propose a novel egocentric representation for visualizing individuals' interaction histories. The egocentric representation conveys two types of roles an individual may play during an interaction process, as an initiator or a responder, with two types of layouts for effectively identifying and comparing interaction patterns.

Data Model and Terminology

We begin our discussion by identifying the key elements and structure of the social interaction process, which provides a basis for the terminology and data model that will be used in our visualization design. (For earlier work in this area, see the "Related Work in Visualizing Time-Oriented Data" sidebar.)

In our day-to-day social experiences, social interactions form the basis of social relations. A *social interaction* can be any relationship between two or more individuals that consists of a sequence of interaction events. It is an essential component that drives various communication technologies. On social media sites such as Twitter, interactions are manifested through tweeting (a user posts a tweet) and replying or retweeting (users rebroadcast a tweet posted by others). Social interactions commonly involve *social objects*, which are the content around which a conversation happens.[3] Examples of social objects include emails (in email exchanges), tweets (in Twitter communications), papers (in coauthorship), and various types of artifacts. A social object connects people with shared interests in a social interaction. There are two types of roles an individual may play during an interaction process: an *initiator* initiates the interaction by creating an social object, and a *responder* responds by acting on the social object created by the initiator. For example, suppose Alice and Bob are two users interacting with each other on Twitter. If Alice posts a tweet and Bob retweets it, then Alice is an initiator, Bob is a responder, and the tweet is a social object.

We introduce a directed tripartite network model to represent the key elements and structure of an interaction process. In Figure 2, initiators and responders are denoted as two types of nodes on either the left or right side, with social objects as the third type of node connecting the initiators and responders. Actions, including initiating and responding to a social object, are denoted as directed edges pointing to the social objects, with a timestamp indicating the time when the action occurred. For example, in Figure 2a, Alice posted three tweets (social objects) at 1:00, 2:00, and 3:30. The first tweet was retweeted by Bob and Carla, the second tweet was retweeted by Carla, and the third tweet was retweeted by Bob and Dan. In this network, Alice is an initiator with actions (posting tweets) represented as red edges,

Related Work in Visualizing Time-Oriented Data

Episogram extends prior work in the visualization of time-oriented data. A summarization of the techniques in this area is available in an earlier study.[1] Here, we compare our work with the most related designs. We focus on comparing our work with the visualizations designed for summarizing social activities in order to understand the design limitations in existing work.

Earlier research efforts have aimed to provide a visual summarization of wide-ranging activities. For example, Michael Ogawa and his colleagues[2] represented the transition of email exchange in open source software projects using Sankey diagrams.[3] Other works employed glyph-based designs to summarize different activities. For example, Robert Erbacher and his colleagues introduced a radial glyph that summarizes a Web server's activity of connecting to other severs over time.[4] The Anemone project introduced a glyph showing the statistical information of users' visiting a webpage.[5] These designs summarized the activities at a given time point as a glyph, and the changes of activities were displayed via animation. PeopleGarden introduced a flower-shaped glyph for summarizing a user's aggregated interaction histories in a discussion group.[6] The different users' flower glyphs are randomly placed in a display area called a garden. Although PeopleGarden summarizes users' interactions, all the details such as when someone was involved in an interaction are unavailable. These designs may be useful in providing snapshot or aggregated views of interaction history, but they are not effective for identifying or comparing temporal patterns in the data. HistoryFlow introduced a stacked flow visualization that displays the collaborations of users who edited the same Wikipedia page.[7] The HistoryFlow visualization lets users compare interaction (coediting a page) patterns within a limited interaction context (a single wiki page). It is thus difficult to extend the design to a more general setting or compare the change of interaction context over time.

References

1. S. Miksch and H. Schumann, *Visualization of Time-Oriented Data*, Springer-Verlag, 2011, chap. 3.
2. M. Ogawa et al., "Visualizing Social Interaction in Open Source Software Projects," *Proc. 6th Int'l Asia-Pacific Symp. Visualization* (APVIS), 2007, pp. 25–32; http://ieeexplore.ieee.org/xpls/abs all.jsp?arnumber=4126214.
3. P. Riehmann, M. Hanfler, and B. Froehlich, "Interactive Sankey Diagrams," *Proc. IEEE Symp. Information Visualization*, 2005, pp. 233–240.
4. R.F. Erbacher, K.L. Walker, and D.A. Frincke, "Intrusion and Misuse Detection in Large-Scale Systems," *IEEE Computer Graphics and Applications*, vol. 22, no. 1, 2002, pp. 38–47.
5. B.J. Fry, "Organic Information Design," master's thesis, *Media Arts & Sciences*, Massachusetts Inst. of Technology, 2000.
6. R. Xiong and J. Donath, "PeopleGarden: Creating Data Portraits for Users," *Proc. ACM Symp. User Interface Software and Technology*, 1999, pp. 37–44.
7. F.B. Viégas, M. Wattenberg, and K. Dave, "Studying Cooperation and Conflict between Authors with History Flow Visualizations," *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2004, pp. 575–582.

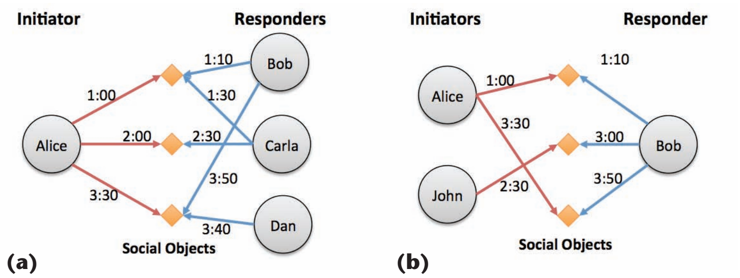

Figure 2. Data model for social interactions. (a) The initiator-centric model and (b) responder-centric model both utilize three types of nodes: initiators, responders, and social objects, which are the gold diamonds connecting the initiators and responders. Actions are denoted as directed edges pointing to the social objects, with a timestamp indicating the time when the action occurred.

the tweets (orange diamond nodes) are social objects, and Bob, Carla, and Dan are three responders whose actions (retweeting) are represented as blue edges. We call this an initiator-centric model because Alice (the initiator) is of central interest for all actions shown in this network.

In contrast, Figure 2b shows a responder-centric model where the responder is of central interest. In this network, Bob (responder) retweeted three tweets posted by Alice and John (initiators). Note that an individual can be both an initiator and a responder at the same time, but in an initiator-centric (responder-centric) model, his/her responding (initiating) actions are omitted.

Social interactions involving a set of initiators and responders can be combined to emphasize the temporal relationship of the interaction events. As Figure 3 shows, interaction events begun by the initiators, Alice and John, are carried on the primary timelines. Each of the social objects created through these initiating actions can be acted upon by different responders. We call the initiating event and the subsequent responding events associated with the same object an *activity thread*. The subsequent responding events on an activity thread are carried on a secondary timeline (as opposed to a primary timeline due to its dependency on the thread's creation). For example, Alice posted three tweets that are starting points of three activity threads. The retweeting events (Dan and Bob's retweeting of the third tweet) are carried on the secondary timelines associated with each of the threads.

An initiator (Alice in Figure 3) can generate multiple activity threads by creating different social objects, and a responder (Bob in Figure 3) can connect to multiple threads by responding to different social objects created by the same or different initiators. Therefore, the problem of visualizing social interaction history can be approached by creating a tool for exploring the various kinds of temporal relationships contained in the connected tripartite networks shown in Figure 3.

Visualizing Social Interaction Data

As we have explained, Episogram helps users explore and compare social interaction data using egocentric viewpoints. The example in Figure 1 shows the social interactions of three physics scholars along various timelines. The scholar (1) F.A. Wilczek has constantly published since early 1970, and most of his papers were published and cited in the journals *Physical Review D* and *Physical Review Letters* (differentiated by color). The scholar (2) James D. Bjorken was productive between 1965 and 1990. His renowned work began with publications in *Physical Review* and later he published more in *Physical Review D*. The scholar (3) Kenneth G. Wilson has an interesting trajectory. He had two productive periods, 1970 to 1975 and 1990 to 1995, and his most-cited work was published in *Reviews of Modern Physics* in 1983. The visualization is generated based on a complete collection of papers published by *Physical Review* as well as citations among them. By visualizing the history of scholars' social interactions with the event contexts (such as journals), this design lets users compare the productivity and impact of these three scholars.

Design Goals and Tasks

The overall goal of our visualization design is to help users gain insights from the social interaction data via data exploration. We have decomposed this goal into a set of tasks that users might seek to answer. We extended the Andrienko task model[2] by characterizing three levels of user tasks in seeking information in social interaction data: elementary, synoptic, and higher-level synoptic tasks.

Elementary tasks address individual data elements. In the context of visualizing interaction history, the user tasks include the following:

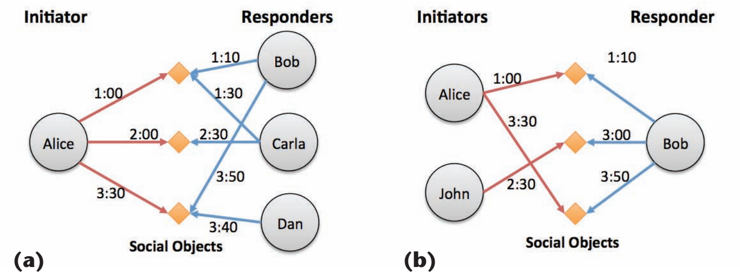

Figure 3. Timeline representation of the interaction model. Interaction events begun by the initiators, Alice and John, are carried on the primary timelines. The subsequent responding events on an activity thread are carried on a secondary timeline.

- *T1 (look up)*: How (through what social object) did actor A interact with actor B at time T (direct lookup)? When did actor A interact with actor B (inverse lookup)?
- *T2 (comparison and relation seeking)*: Compare how actor A interacted with actor B as opposed to with actor C (direct comparison). When actor A initiated an interaction by creating a social object, did actor B respond before or after others (inverse comparison)? When did actor B respond to actor A quicker than others (relation seeking)?

Synoptic tasks involve a general view of data. Here, the user tasks include the following:

- *T3 (pattern identification and search)*: What was the frequency of interaction between actor A and others during time T (pattern identification)? When did actor A interact with others frequently (pattern search)?
- *T4 (pattern comparison)*: Compare the interaction frequency between actor A and others during time T_1 and time T_2. How do others respond to actor A during T_1 and T_2?

One of the key motivations for visualizing social interaction data is to characterize individuals' social behavior and further gain insights by exploring how people's interactions with others might affect their life outcomes (for example, work productivity or career path). Hence, we identify *higher-level synoptic tasks* (that is, more abstract tasks) based on the identification, search, and comparison of patterns about individual social actors.

- *T5 (actor pattern identification and search)*: Did actor A's interactions with others persist over a long time period, or were they concentrated during a certain time? When did actor A's interactions with others suddenly increase?
- *T6 (actor pattern comparison)*: How did actor A's interactions with others different from those of actor B? Was actor A more active (in terms of

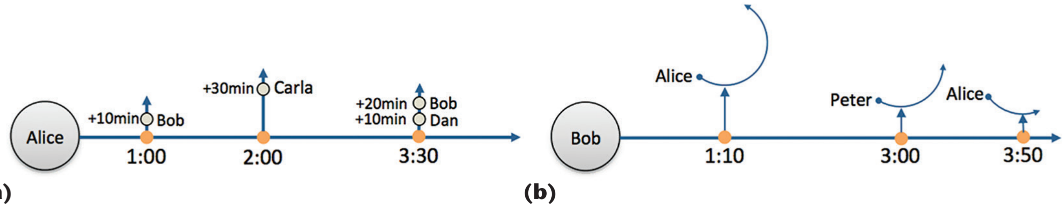

Figure 4. Episogram design overview based on the combination of two different views: (a) initiator view and (b) responder view. These views were extracted from the networks in Figure 3. The vertical lines indicate the activity threads, and the circles indicate the corresponding social object. The crescent-shape curved arrows denote interaction threads.

initiating an interaction) than actor B? Was actor A more responsive (in terms of responding to others' interactions) than actor B?

We designed Episogram iteratively based on these tasks by working closely with an expert with a background in computational social science. We held weekly discussions for approximately 1.5 months in order to develop an effective visual design. In each design iteration, we proposed and manually drew several design choices based on a small set of toy data for illustrating the concept. The expert evaluated their effectiveness, identified their limitations, and provided design suggestions for improvements by applying them to solve the aforementioned tasks.

Ultimately, two designs (a Gantt chart and the Episogram design proposed here) were considered to be the most effective among all other design choices. We conducted a formal, controlled user study (which we describe later) to compare these two designs. The results illustrated several significant benefits of the Episogram design.

Visualization Design

Our Episogram design seeks to help users find answers for tasks T1–T6 from the temporal social interaction data as illustrated in Figure 3. We propose an egocentric representation to focus on each individual's interaction at a given time, based on the role he or she plays in the social interactions. In particular, this egocentric data can be shown in the initiator view (Figure 4a), which shows when and how the individual initiated interaction events by creating social objects, and in the responder view (Figure 4b), which shows when and how the individual participated in interactions with respect to social objects created by others. The two views in Figures 4a and 4b were extracted from the networks in Figure 3.

In Figure 4a, the primary timeline shows time points when Alice posted tweets. Each activity thread, represented by the vertical line in Figure 5a, interacts with the primary timeline at the time point t, which is the time when the corresponding social object (shown as a circle at the intersection) is created. All subsequent responding events with respect to the social object are marked on the vertical line with intersections indicating when the responding events occurred. The length of the vertical line depends on the lag of the last responding events on the thread.

In the responder view shown in Figure 4b, the primary timeline carries time points when Bob retweeted others' tweets. Each interaction thread in Figure 5b is represented as a crescent-shape curved arrow upheld by a vertical line. The crescent shape begins with a circle (representing the corresponding social object) indicating the thread's creation—the time when the social object is created. The crescent shape ranges from 0 to 180 degrees, indicating the relative duration of the corresponding social interaction thread. A 180-degree crescent shape represents the longest duration of the activity thread in the dataset. The length of the vertical line double encodes the duration of the responded thread. The intersection between the crescent shape and the vertical line shows when the responder participated in the activity thread—for example, the responder's retweeting time. Hence, the orientation of the crescent shape reflects how early or late a responder participates in the activity thread.

In both views, the thread color and size can be used to represent additional data attributes such as the sentiment and the number of retweets. In addition, connecting the vertical thread lines to the primary timeline and arranging them parallel to their start points facilitates a fast comparison of different thread durations, thus enabling an easy detection of influential threads.

Thread Aggregation

We developed a thread aggregation design to reduce the visual clutter caused by dense social interaction events and help users detect potential events in social interactions.

In the initiator view, a cluster of threads can be visualized by directly merging multiple threads into the same vertical line. This shared vertical line starts at the time of the earliest created social object, records the time points of all responding events with respect to all social objects included in this thread cluster, and ends at the time of the last responding event. Figure 6a shows an example of a thread cluster that includes the threads shown in Figure 4a.

In the responder view, we visualize the thread cluster by adding curved lines inside a crescent shape. The arc of the crescent shape represents the overall time span of all activity threads included in this cluster, and each of the curved lines represents how the particular thread spans relative to the overall time span. The vertical line is attached with horizontal arms that point to the time points when the responder responds to the corresponding threads included in this thread cluster. The *y* position of each arm is determined by height of the vertical line of the corresponding thread, showing its duration. Figure 6b shows an example of a thread cluster that includes the threads shown in Figure 4b.

To detect events, we cluster activity threads by using mean shift,[4] a nonparametric analysis technique that adaptively generates clusters that are always centered at the positions with the highest densities in the data space. We select thread features for clustering by considering the threads' closeness on the primary timeline and their semantic similarities in content (for example, the tweets' topics).

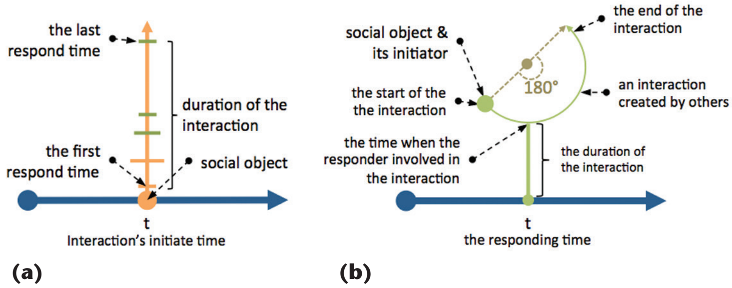

Figure 5. Episogram activity thread: (a) initiator view and (b) responder view. The crescent's shape (which can range from 0 to 180 degrees) indicates the relative duration of the corresponding social interaction thread. The intersection between the crescent shape and the vertical line shows when the responder participated in the activity thread.

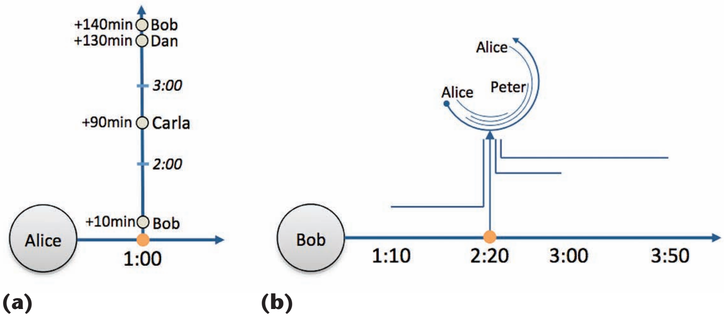

Figure 6. Visualizing a thread cluster in (a) initiator and (b) responder views. These two examples show the aggregation of the activity threads in Figures 4a and 4b, respectively.

System Interface and Interactions

We implemented Episogram as a Web application. The system interface (see Figure 1) consists of four components: a toolbar, the main display, a legend, and an actor list, which correspond with the (a) to (d) labels in Figure 1, respectively.

With these components, users can select different datasets using a dropdown menu and select one or more actors to be visualized from the actor list. With the toolbar, users can select different views (initiator versus responder) to visualize the selected actors. Also, when the data are densely distributed over time, users can aggregate the threads using automatic event detection, selection, or the categorical attributes associated with the corresponding social objects. Lastly, users can zoom into a particular time period by selecting a range on the time axis shown at the top of the main display, or they can select a thread to focus on by clicking on it. The focused thread will be highlighted while the others will appear in grey.

Case Studies

To illustrate how our design can be used to explore and identify patterns in social interaction data, we use two datasets that capture social interactions in different contexts. The first dataset consists of Twitter users' interactions (posted tweets and retweets) around political debates, and the second dataset consists of academic publications in physics journals that capture scholars' interactions in terms of publishing and citing papers.

Detecting Anomalous Behaviors in Twitter

The Twitter dataset we used was collected during the US presidential election debates held in October 2012.[5] For demonstration purposes, we selected a set of the most active users who posted or retweeted the most in the data.

Figure 7 shows the initiator view of three selected users with different posting behaviors. Most users in the dataset exhibit scattered events similar to the user SirrK (Figure 7a), whose tweets were posted at different times across the data period and some of the tweets received more retweets than others. In Figure 7a, the activity threads are colored based on the sentiments of the corresponding tweets (red is negative, yellow neutral, and green positive).

The users Perciousliberty and ObamaVSAmerica exhibit different patterns from those of typical users. In particular, Figure 7b shows that Perciousliberty posted large numbers of tweets regularly at a particular time period each day. Figure 7c shows that ObamaVSAmerica continuously posted enormous numbers of tweets throughout the entire

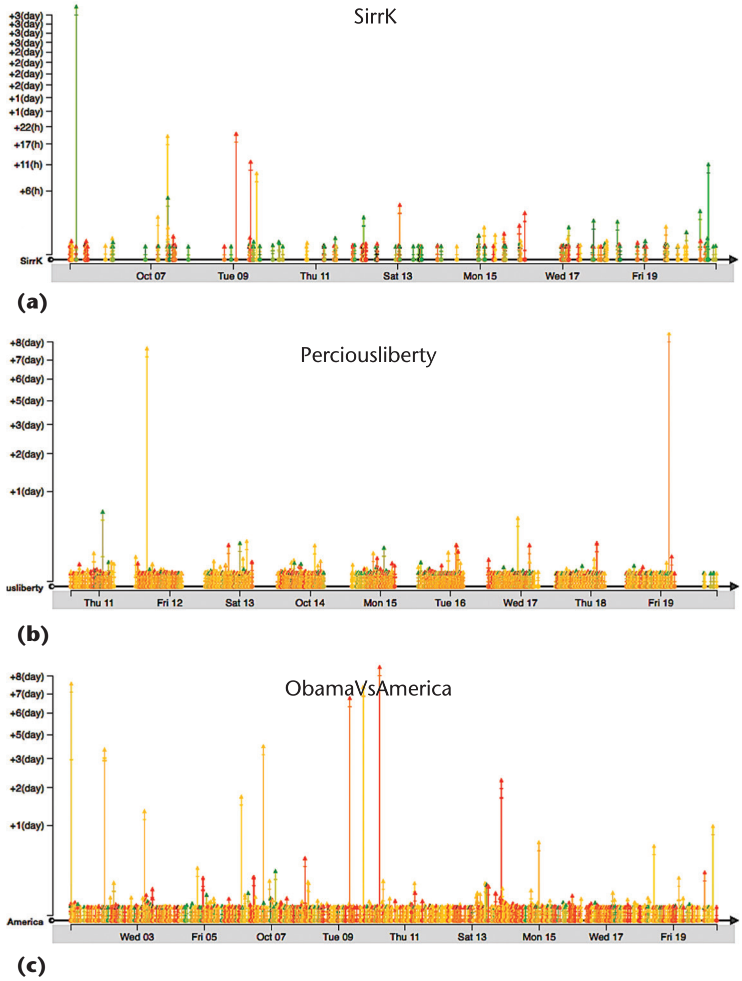

Figure 7. Summarizing and comparing Twitter users' posting behaviors in Episogram using the initiator view. The graphs illustrate (a) typical posting behavior, (b) periodical posting behavior, and (c) a continuous posting behavior. The colors denote the sentiments of the corresponding tweets, where red is negative, yellow neutral, and green positive.

data period. Negative sentiments are pervasive in these tweets, which can be observed in the red-colored activity threads. By reading the content of the tweets posted by the two users, we found that most of these tweets express sentiments against the Obama administration. Interestingly, such strong and persistent "attacks" in Twitter communication can be easily identified by visualizing the temporal patterns of posting events.

Figure 8 shows the responder view for two different users. The primary timeline records the time when the selected user retweeted other user's tweets, and the activity threads show how early or late the selected user's retweeting time compared with other retweeting users' with respect to the same tweets. Figure 8b shows that the user CWade91 tended to retweet others' tweets immediately after the tweets were posted; the vertical lines of these activity threads intersect mostly with the beginning of the crescent shapes. This feature can be identified more clearly by using the thread aggregation function (Figure 8c), which displays clusters of threads when the retweeting events occur close in time. This early retweeting tendency suggests that the user is an active information spreader in Twitter. In comparison, Figure 8a shows that the user JsrRoger exhibits a more typical responding pattern; his/her retweeting of a tweet of interest may be earlier or later than other users for the same tweet.

Visualizing Researchers' Career Path

The publication dataset is a complete collection of papers published in *Physical Review* as well as citations among them. It covers papers published in different journals such as *Physical Review* (*PR*), *Physical Review Letters* (*PRL*), *Reviews of Modern Physics* (*RMP*), and *Physical Review A*, *B*, *C*, *D*, and *E*, each of which focuses on a specific area in physics. As exemplary cases, we selected a set of scientists who are mostly Nobel laureates or major prize/medal awardees. All their papers, references, and citations are included for demonstration purposes.

When visualizing publication data in Episogram, the initiator view illustrates a researcher's productivity over time as well as his/her research impact generated by these publications. Each thread centers around a paper published by the researcher, indicating how the paper was cited by others over time. The responder view, on the other hand, visualizes the way in which the papers by this researcher cited existing studies. Each thread shows, in an aggregated fashion, how a paper by the researcher cited other existing papers. Each of the cited papers is represented as an arc in the aggregated thread. In both views, the threads are colored by the journals in which the threads' corresponding papers were published. Using this encoding scheme, we demonstrate the Episogram's power of interpreting a researcher's career path.

In our first example, we take H. Eugene Stanley as an exemplary case for our study. He is an American physicist who has made many seminal contributions to several topics in statistical physics and was awarded the Boltzmann Medal for his contributions to phase transitions.

A first glance at Stanley's career illustrated in Figure 9a leaves two impressions. First, the high intensity of vertical bars over time make it immediately clear that Stanley has been highly productive throughout his career. Second, his publications as well as citations to these publications are characterized by a mix of different colors: blue corresponds to papers in premier physics journals that cover all areas of physics (*PR* is in dark blue and *PRL* in light blue), whereas green and red correspond to journals specializing in a particular physics domain (*PRB* in green covers condensed matter physics, *PRE* in pink covers statistical physics and interdisciplinary physics, and *PRA* in red covers atomic, molecular, and optical physics). Hence, the mix of blue with other colors indicates publications in both premier journals that are of interest to different physics domains and papers specializing in a particular field. We also observe a general shift in color from green to red/pink over time, documenting changes in research topics along his career.

More precisely, at the beginning of Stanley's career, he published most of his papers in *PRL*, a high-impact premier journal that covers all physics topics. The primary color (green) of the citations to these papers indicates their fundamental impact on condensed matter physics. From 1971 to 1976, Stanley was extremely productive, and the high intensity of green bars during the period indicates extensive publications by him on condensed matter physics (see Figure 9b). The height of these bars indicates the high impact of these papers. The dense horizontal green bars in each thread signal that his papers made significant advances within the research field. The next two decades following this significant burst of publications mark a gradual shift in his research focus. With colors shifting from green to red (see Figure 9c), Episogram demonstrates an increasing focus on atomic and molecular physics as well as statistical physics in his research agenda. During this period, his publications represent a great mix of papers in light blue together with green and red. Such a mix indicates that his research covers both papers in *PRL* that are general to all areas of physics and require more rapid dissemination and more detailed papers that impact a specific domain.

Episogram also reflects historical changes in scientific publications. From 1990 to 1993, there was a gradual split of *PRA* into two journals, *PRA* and *PRE*, with *PRE* focusing on statistical physics, plasmas, fluids, and related interdisciplinary topics. Clearly, Stanley's research is related to the focus of *PRE*, as we observe an interesting change of colors from red to pink following this journal split. In addition, we can see a general decrease in the height of vertical bars, as more recent papers have less time to accumulate their citations.

Figure 10a shows the aggregated responder view of the same data, providing us with another perspective on Stanley's career based on the way he references other papers. In the early stage of his career, he mostly cited the latest papers in his publications, showing he was as an early adopter of new ideas, which partially explains the observed

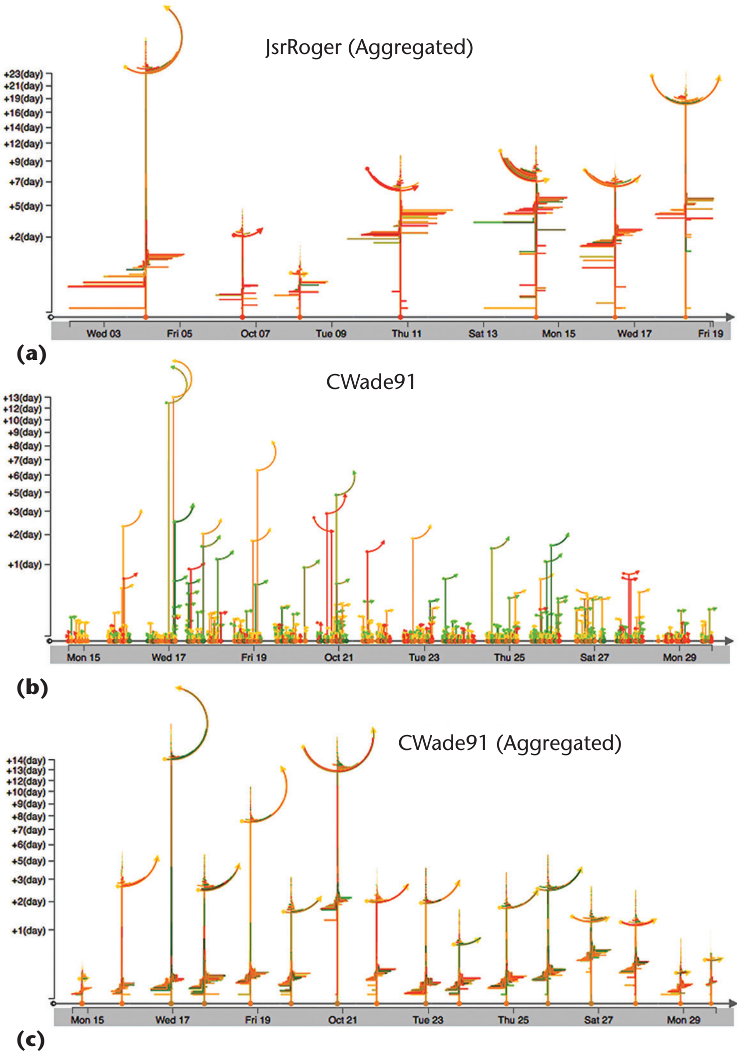

Figure 8. Summarizing and comparing Twitter users' retweeting behaviors in Episogram using the responder view. The graphs illustrate (a) a typical retweeting behavior, (b) monitoring behavior, and (c) the aggregation view of the monitoring behavior.

impact of his work. In later stages, especially after 1995, he cited a higher fraction of older papers in his publications. This pattern is potentially due to a combination of two factors, including the temporal cutoff of our dataset in 2009 and his increasing focus on well-known or longstanding problems in his research. The aggregated threads computed by mean shift bring more visual clarity for the observed patterns (see Figure 10b).

User Study and Discussion

We conducted a controlled within-subject study to compare Episogram with the traditional timeline view, Gantt chart, based on a set of pattern exploration tasks. (For details about the study and interviews, see supplemental materials at http://nancao.org/pubs/cao_cga_episogram_si.pdf.) The study results suggested the design effectively conveyed both detailed and overall pictures of different actors'

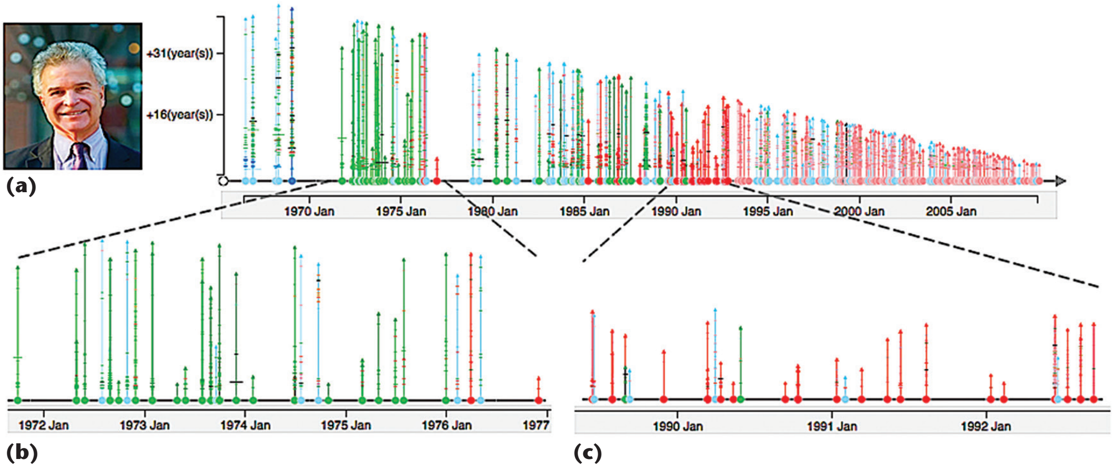

Figure 9. Summarizing H. Eugene Stanley's career path in the initiator view: (a) overview of publication records, (b) the time period in which the professor was productive in condensed matter physics, and (c) the time period in which the professor focused on atomic, molecular, and optical physics. The colors indicate the various publication journals: blue corresponds to papers in premier physics journals that cover all areas of physics, whereas green and red correspond to journals specializing in a particular physics domain.

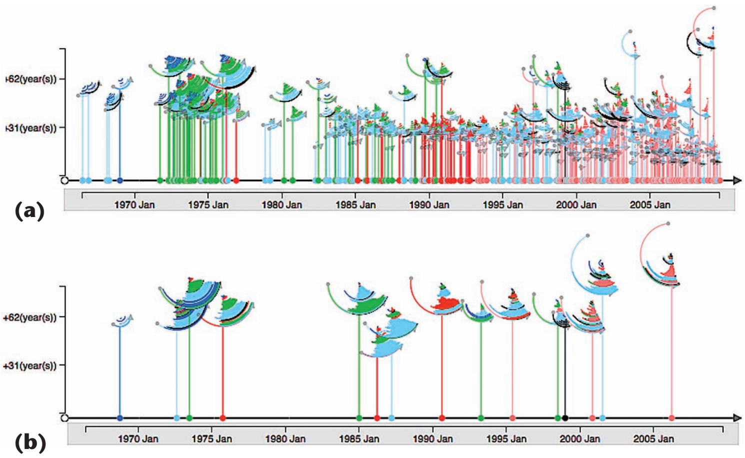

Figure 10. Visualization of H. Eugene Stanley's publications in responder view: (a) threads of individual publications and (b) thread aggregation.

social interactions by helping users identify data elements in the elementary tasks, identify interaction patterns in the synoptic tasks, and characterize actor interaction tendency in the higher-level synoptic tasks.

Particularly, our design complements the existing network representations by offering users a summarization of the interaction history that facilitates an understanding of how individual actors act and react as part of a larger network. When compared with traditional timeline views, Episogram has many key features. The timeline view resembles a typical compound-event-based timeline design such as the Gantt chart, in which the primary timeline and the activity threads share a common time axis. However, in this layout, the activity threads and primary timeline may be displayed far apart when data increases, making it difficult to identify and compare patterns. Episogram, on the other hand, directly connects the primary timeline with activity threads, clearly providing the context of an interaction event and its subsequent events. Each thread is displayed with the length encoding the thread's relative duration. Despite its limitation of conveying exact temporal information, the design decision was made to allow users to easily compare the temporal relationship of the interaction event initiated by or responded to by the actors of interest.

In addition to the controlled user study, we also interviewed two expert users from different but related disciplines. The first expert is a PhD candidate in applied mathematics and computer science from a European university with expertise in social networks and human mobility. The second expert is a postdoctoral fellow in physics from the United States with expertise in network science. Both experts have published extensively on social network analysis, and they are familiar with the publication datasets used in our study. Both experts were impressed by the rich information offered by Episogram as well as the design itself. The first expert particularly appreciated that Episogram translates the citation statistics into visual patterns: "First time you get to look at these patterns!" The second expert highlighted the utility of our tool by comparing it with the simple or aggregated charts provided in citation search engines such as Google Scholar (http://scholar.google.com). She pointed out that one novel aspect of our tool is that it allows users to see how a scholar was cited by others in the absolute and relative temporal dimensions, and thus we "have all scientists' productivity at a glance." Both experts agreed independently that the most useful and interesting feature offered by our tool is the aggregation function—that is, papers or citations can be aggregated by similarity and still differentiated by their published journals. They believe this "is a useful approach for reducing the clutter."

Based on our studies and interviews, we also note that our design has some limitations mentioned by our users and experts. First, the egocentric design does not let users view all interactions between any two actors in a social network. We believe this limitation can be addressed by integrating our current design with a typical node-link network representation. The second limitation concerns overplotting: the rich patterns provided in the activity threads can be overwhelming if the selected actor was particularly active or productive. In a real-world dataset, the chance of seeing the cluttered activities for an actor is rare due to the well-known power-law phenomena.[6] However, when users are interested in visualizing actors with many activities, there are several ways to effectively reduce the visual clutter: users can select activities by categorical attributes, zoom in to a particular time period, and aggregate activities using the aggregation function. We believe these additional tools help balance the richness and clarity in our original visual design.

Episogram is an interactive visualization for exploring and summarizing social interaction data. Our design aims to assist in a variety of user tasks ranging from elementary tasks to higher-level pattern discovery. It allows users to generate multiple views for different actors' social interaction history and compare multiple actors in an integrated display. Our evaluation, including case studies and a controlled user study, have demonstrated its usefulness.

Our future work includes two directions. First, we plan to conduct user studies to evaluate the scalability of our visual designs. Second, we intend to develop visual analysis systems for detecting, analyzing, and visualizing different user behaviors via Episogram and other types of visualizations such as node-link graphs. We will also apply this system to analyze other datasets such as email archives.

Acknowledgments

This work is partially sponsored by DARPA under the Social Media in Strategic Communication (SMISC) program (agreement W911NF-12-C-0028). Dashun Wang is supported by the US Air Force Office of Scientific Research under agreement number FA9550-15-1-0162. We thank P. Deville and R. Sinatra for providing us with author disambiguated datasets as well as many useful discussions along the way.

References

1. P. Dourish and V. Bellotti, "Awareness and Coordination in Shared Workspaces," *Proc. ACM Conf. Computer-Supported Cooperative Work*, 1992, pp. 107–114; http://dl.acm.org/citation.cfm?id=143468.
2. S. Miksch and H. Schumann, *Visualization of Time-Oriented Data*, Springer-Verlag, 2011, chap. 3.
3. N. Simon, *The Participatory Museum*, Museum 2.0, 2010.
4. Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, 1995, pp. 790–799.
5. Y.-R. Lin et al., "Rising Tides or Rising Stars? Dynamics of Shared Attention on Twitter during Media Events," *PloS One*, vol. 9, no. 5, 2014; http://dx.doi.org/10.1371/journal.pone.0094093.
6. M.E. Newman, "The Structure and Function of Complex Networks," *SIAM Rev.*, vol. 45, no. 2, 2003, pp. 167–256; http://epubs.siam.org/doi/abs/10.1137/S003614450342480.

Nan Cao *is a full professor in College of Design and Innovation at TongJi University and an adjunct professor of computer science at NYU ShangHai and NYU Tandon. His research interests include data visualization, visual analysis, and data science. Cao has a PhD in computer science from the Hong Kong University of Science and Technology. Contact him at nan.cao@gmail.com.*

Yu-Ru Lin *is an assistant professor in the School of Information Sciences at the University of Pittsburgh. Her research interests include human mobility, social and political network dynamics, and computational social science. Lin has a PhD in computer science from Arizona State University. Contact her at yurulin@pitt.edu.*

Fan Du *is a PhD student in computer science at the University of Maryland, College Park. His research focuses on data visualization and human-computer interaction, especially on analyzing healthcare data and user activity logs. Du has a bachelor's degree in computer science from Zhejiang University. Contact him at fan@cs.umd.edu.*

Dashun Wang *is an associate professor in the Kellogg School of Management at Northwestern University. Through the lens of large-scale datasets, his work focuses on using and developing tools of network science to help improve the way in which we understand complexity and discover the underlying principles governing self-organized systems. Wang has a PhD in physics from Northeastern University. Contact him at dashun.wang@kellogg.northwestern.edu.*