A Multiscale Approach to Network Event Identification using Geolocated Twitter Data

Chao Yang
School of Computing
University of Utah
chaoyang@cs.utah.edu

Ian Jensen
School of Computing
University of Utah
ian.jensen@utah.edu

Paul Rosen
SCI Institute
University of Utah
prosen@sci.utah.edu

Abstract

The large volume of data associated with social networks hinders the unaided user from interpreting network content in real time. This problem is compounded by the fact that there are limited tools available for enabling robust visual social network exploration. We present a network activity visualization using a novel aggregation glyph called the *clyph*. The clyph intuitively combines spatial, temporal, and quantity data about multiple network events. We also present several case studies where major network events were easily identified using clyphs, establishing them as a powerful aid for network users and owners.

1 Introduction

Given the volume of data available on modern social networks, it has become difficult for an unaided user to query and interpret the content of the network in a near real-time environment. By arming the user with an intuitive visual analysis tool for twitter, we aim to empower individuals, such as network users or owners, to better understand the characteristics of social network activities in the context of time, location, and content.

Direct representation of this data does not work well since there are spatial, temporal, and textual components to a tweet. Combining these data elements in any naive fashion will produce a visualization which is too cluttered to be effective, and removing any component will detract from the explanatory power enabled by the visualization. Beyond just including these elements in a visual design, the ability to explore the data at many scales (i.e. statewide, countywide, citywide, etc.) must be incorporated for the visualization to be truly useful.

To address these challenges, we produced a system which ties together the spatial, temporal, and quantity data associated with tweets into single a streamlined visualization with textual data summarized in a linked companion interface. This was accomplished using our novel visual representation of spatiotemporal data dubbed the "clock" glyph or ***clyph***, as shown in Figure 4. The clyph combines the locations of the tweets and the range of times at which they occur for an abstracted area. Clyph benefits include:

1. A visual representation which combines raw tweet data from many tweets while minimizing the data loss inherent in aggregation.
2. A tool for exploring tweet geography at multiple scales, emphasizing trends in time and location.
3. A tool that enables users to detect major network events (i.e. concerts, conventions, etc.).

Figure 1 demonstrates how a user might interact with our system. The clyph is used as a visual abstraction of multiple tweets which are spatially similar but temporally varied. The clyph is located in the center of all tweets it abstracts. The interior of the clyph displays temporal data marking the median, quartiles, and range for tweet times. The notches on the perimeter indicate both the number and relative direction of each tweet abstracted by the clyph. A companion text display is provided which displays a list of tweets for a selected clyph, along with the 20 most frequently used words. The example interaction in Figure 1 shows a user beginning with a citywide view of twitter data for Salt Lake City, then progressively zooming in to further differentiate tweets. At each level of zoom, the clyphs are recalculated to maximize the display of information while avoiding any overlap. As the user explores the city at the top-level zoom, anomalous keywords hint at possible events (Figure 1 left). As the user zooms, the clyphs begin to differentiate from one another and more event-related keywords appear in some for some clyphs (Figure 1 middle). At the lowest-level zoom, the event is localized in both location and context (Figure 1 right). Finally, the clyphs and their tweets reveal that an outdoor retailer's market has taken place near the local convention center.

2 Related Work

Social network data can be explored in numerous ways, the majority which explore relationships within the network.

Relationships within social networks are often explored using node-link diagrams. For example, Heer and Boyd [9] designed an application for visual exploration and analysis of online social networks by using node-link network layouts. There are several well-known techniques for improving the effectiveness of node-link diagrams [19] which use strategies to group the visualization of nodes into regions according to additional attributes such as categorical, ordinal, and binned numerical data in node-link diagrams. Node-link diagrams have also been enhanced by using adjacency matrices as linked-views, providing a hybrid representation that draws heavily upon two traditional node-link diagram representations. Finally, Brandes and Kick [2] presented a gestaltline approach that shows type, extent, and time of relationships.

When graph data is attached to geospatial data, such as that of Twitter follower networks, flow mapping has

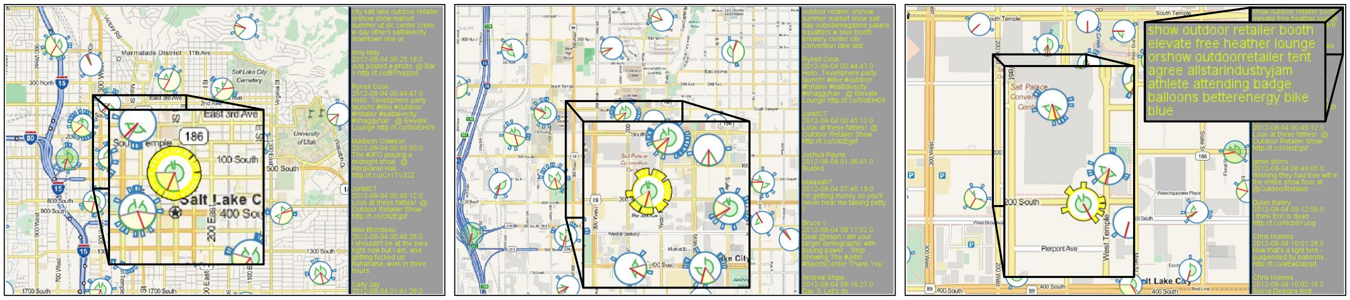


Figure 1: Example interaction at three scales. Left: Highest zoom gives an overview of the city. Middle: Midscale visualization begins to show more detailed information. Right: Tightest zoom level points towards specific events.

been an effective visualization tool. Rae [17] applied flow mapping within contemporary GIS by mapping a large migration matrix from the United Kingdom's 2001 census. Guo [6] used several methods such as hierarchical regionalization, flow mapping, and multivariate clustering and visualization to discover major flow patterns relations from migration data in the United States. This graph data is often noisy. A number of approaches such as edge clustering methods [16, 4] and edge bundling methods [10] have been used to reduce the noise, helping generate flow maps.

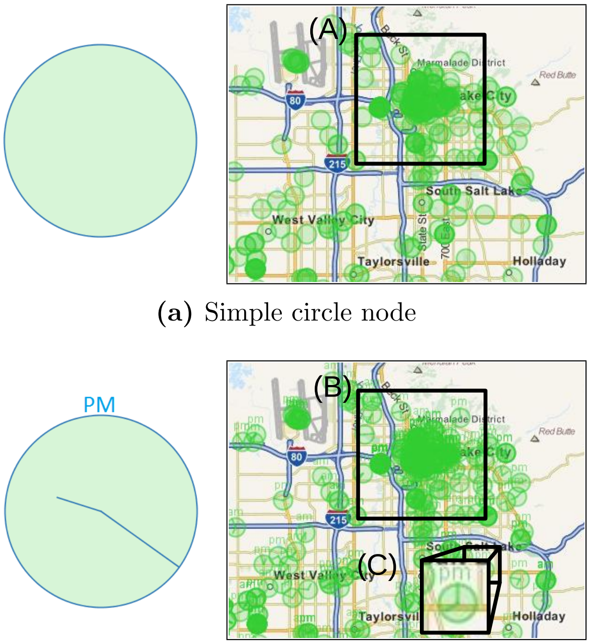
Perer and Shneiderman [15] proposed a system that uses attribute ranking and coordinated views to allow users to explore social networks using overviews, filtering nodes, finding outliers, and visually coding the network visualization. This interactive application inspired further research in social network visualization. Groh et al. [5] proposed a dynamic social network visualizer to visualize temporal social network data. It introduced the 3D interpolated NURBS "tubes" to represent activity and social proximity for a certain actor. It is one of the few attempts in applying 3D to social network visualization. Luo et al. [13] introduced a spatial-social network visualization tool, the GeoSocialApp, which provided the geographical, network, and attribute views to help explore the different attributes of spatial-social network data. This system has been extremely informative in our research, as it encodes datasets similar to ours and emphasizes dimensions we are also interested in. Cho et al. [3] explored patterns of human mobility on three large datasets using statistical methods and visualization methods. The network flip books and dynamic movies introduced by McFarland [11] provided insight into some of the interactive aspects of a network visualization.

Several methods are proposed to solve the problem of effectively visualizing the multivariate and multidimensional data. Guo et al. introduced computational and geographic methods to explore and visualize multivariate spatial patterns within high-dimensional geographic data [8] and later derived complex patterns from spatiotemporal and multivariate data sets [7]. While our data is not necessarily as high-dimensional as what is discussed in these papers, they did inform the design of our own system. In addition to visualization solutions for multidimensional data, we also briefly explored using simple data mining techniques to extract meaning from the data. Keim and Kriege [12] evaluated their own visual data mining techniques and compare them to other popular techniques for visualizing multidimensional data.

3 Visualizing Individual Tweets

In order to identify network events within tweet data, we began by exploring methods to visualize individual tweets simultaneously.

We began with the most obvious abstraction for spatial data, a simple circular glyph representing a single tweet (Figure 2a). Placing many of the circular glyphs on a map allowed spatial clusters to become very apparent, as seen in Figure 2, box A. However, the visualization did not show any temporal references or temporal clustering. This could cause misinterpretation of a busy place as interesting event. Animation, i.e. glyphs appearing and disappearing,



(a) Simple circle node

(b) Individual clock glyph

Figure 2: Individual tweet visualizations

was used as an alternative, but the length of tweet visibility limited the bandwidth of event length. For example, short visibility meant the long-duration events were lost, while long visibility would obscure short-duration events.

Our next approach was to enhance the interface with temporal information. We settled upon commonly understood design metaphor, a wall clock. Each circular glyph had a minute and hour hand added to its display, along with text indicating a.m. or p.m. (Figure 2b). For sparsely located, non-overlapping tweets, this representation gives quick access to both the location and timing of the tweet, as seen in Figure 2, box C. However, as tweets begin to gather spatially, visual cluster ensued, and the information was lost, as seen in Figure 2, box B.

4 Tweet Aggregation

The number of tweets presented in the display can be quite large, in particular when large areas or time spans are cover. To effectively display all of this data, it is necessary to aggregate raw the data so that it can be represented by a glyph which minimizes or eliminates any overlap between neighbors. We did this by clustering tweet with nearby locations. However, finding the combination of glyphs is an optimization problem which is computationally expensive. In order to maintain interactive exploration, we chose to use a greedy method, based upon the node-grouping algorithm proposed by Newman [14], for selecting the location of glyphs which completes in worst case $O(n^2)$ time.

```
  Data: List $\mathbb{M}$ which contain all tweet locations
        within the current range (as defined by the
        zoom level and location)
  Result: Set $\mathbb{P}$ of glyph locations
1 while $\mathbb{M}$ is not empty do
2     Let $p$ be a random point from $\mathbb{M}$
3     Insert $p$ into $\mathbb{P}$
4     foreach $m$ in $\mathbb{M}$ do
5         if $distance(p, m) < 2r$ then
6             Remove $m$ from $\mathbb{M}$
7         end
8     end
9 end
```

Algorithm 1: Tweet clustering algorithm

Our aggregation algorithm, described in Algorithm 1, takes input $\mathbb{M}$, the list of all tweet locations for the current configuration. A point p is then randomly chosen from $\mathbb{M}$ (line 2). All points in $\mathbb{M}$ within radius $2r$ (twice the radius of the glyph) of p are removed from $\mathbb{M}$ and assigned to the glyph centered at p (lines 4-6). The selection of a radius of $2r$ leaves our layout somewhat sparse, but guarantees that no two glyphs will overlap. This process is repeated until all points are assigned.

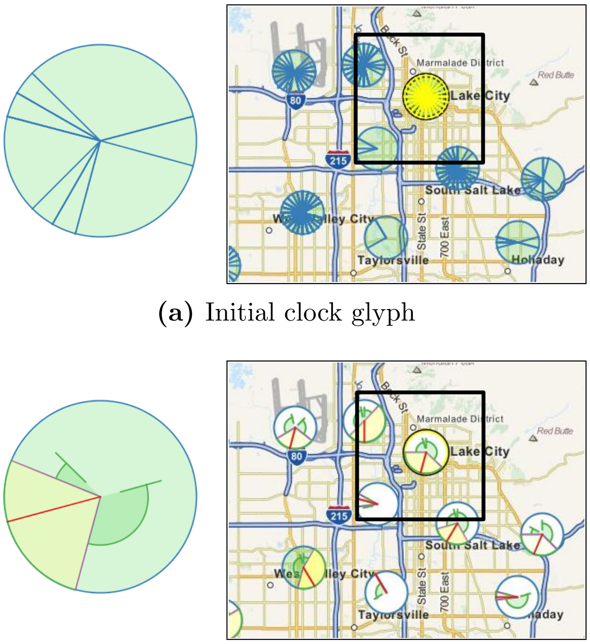
In addition to spatial aggregation, we also enable the user to adjust the temporal range of the data. By providing a time slider, we facilitate identifying events around times of interest in addition to locations of interest through the map view.

5 Visualizing Multiple Tweets with Clyphs

Our spatiotemporal aggregation effectively removed the clutter that previously plagued our visual interface; however, significant information is lost to aggregation. Therefore, our glyph needed further refinement improve the volume of information communicated.

Our next approach was to include timing information for multiple tweets within a single glyph. This was done by extending the clock representation to a single handed clock. Now, the clock glyph can represent multiple tweets by placing one mark for each tweet within the glyph representing it (Figure 3a). In addition to the advantage of reducing clutter, this design also implicitly encodes of the volume of tweet activity. However, as the number of tweets and thereby the number of marks grow, this interface can become cluttered. If too many lines are drawn, they are no longer differentiable. This will limit the ability to visually measure quantity. Even worse, the clutter can lead to misinterpretation of quantity. For example, the boxed glyph in Figure 3a shows dense twitter activity at apparently all hours of the day. In this case, no further meaning can be derived.

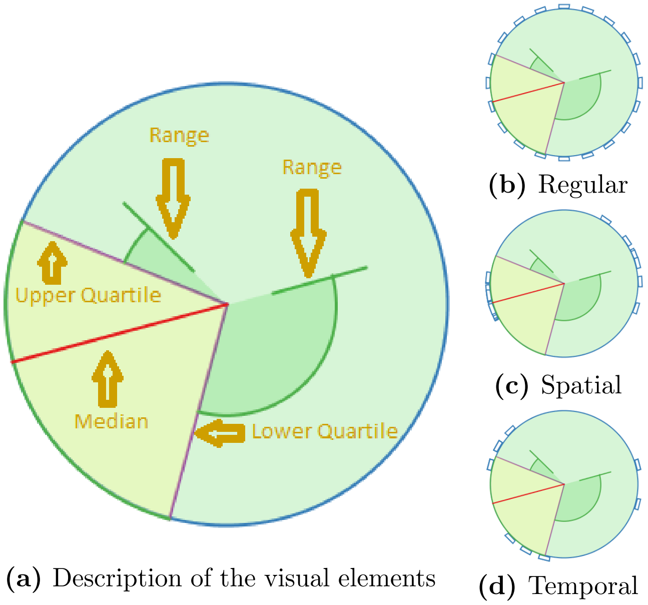
To prevent this potential clutter, we decided to maintain our clock metaphor, but move to a statistical view of the data. We decided upon median, upper and lower quartile, and upper and lower range as the most significant statistical elements of the data. Figure 4a demonstrates



(a) Initial clock glyph

(b) Quartile glyph

Figure 3: Multiple tweet visualizations



(a) Description of the visual elements **(b)** Regular **(c)** Spatial **(d)** Temporal

Figure 4: Description of visual elements (left) and variations of notch location (right) for the clyph

how we integrated these elements into the clyph. The median value is represented by a solid red line. The region between the upper and lower quartiles is colored in a yellow-orange tone. Finally, the range is represented by the region marked in green. Figure 3b shows this version of the glyph in action. In Figure 3a, meaning was lost with too many marks. However, in Figure 3b, the quartiles and median of the data in the glyph tell a different story. Inspection of the glyph shows that the bulk of the twitter activity in this region occurred between approximately 10:00 and 18:00. This difference makes it apparent that aggregating the noisy data into the glyph enables the user to better derive meaning from the data.

This statistical view allows the representation to scale to any quantity of tweet data without cluttering the display. However, this new presentation loses any sense of quantity of tweets. The final clyph representation, presented in Figure 4, has marks distributed along the outside of the clyph, each one representing a different tweet. This design represents the quantity of tweets as well as the distribution. We present three alternative to the placement of the notches. The first (Figure 4b), places notches at evenly distributed locations, giving only a sense of volume. The second (Figure 4c), places notches at the spatial direction of the tweet, relative to the centroid of the clyph. This gives access to additional spatial information. The final version (Figure 4d) places notches at the time in which each tweet occurred giving a better sense of temporal distribution.

6 Implementation

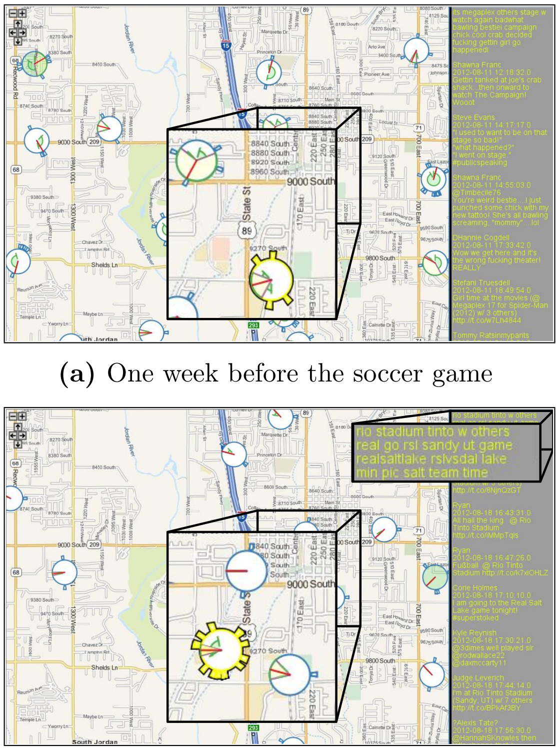
The data used in our experiments was obtained using a crawler we developed. The Java-based crawler collects a stream of tweet data and stores it into a MySQL database. Data collection was limited to only geolocated tweets within the state of Utah, though our examples focus on Salt Lake City and surrounding areas. The data was collected from August 1, 2012 through August 26, 2012. Since we only collect tweets with location information, our results were limited to approximately 184,000 total tweets or slightly over 7,000 individual tweets per day. The database collected over that time frame was 195 MB.

Our visualization tool was written in Java using Processing [18] and features an interactive interface which renders at 25 frames per second. The visualization tool directly queries our MySQL database for near real-time updating of visualization results.

The interface displays a map, provided by modest maps [1], and places clyphs at appropriate locations. A time slider enables the user to pick the time frame for which the clyphs are generated. A side panel provides textual feedback, listing all tweets from a selected clyph. In addition, the 20 most frequently used words (excluding common words) are extracted from the tweets of the selected clyph. We found these keywords exceedingly useful in determining the purpose of an event (i.e. fair, concert, sporting event, etc.) after it was located using our visualization.

7 Case Studies

We now present a series of case studies which verify our tool's usefulness at identifying large network events.



its megaplex others stage w watch again badwhat bawling bestiei campaign chick cool crab decided fucking gettin girl go happenedi

Shawna Franc
2012-08-11 12:18:32.0
Gettin tanked at joe's crab shack...then onward to watch The Campaign! Wooot

Steve Evans
2012-08-11 14:17:17.0
"I used to want to be on that stage so bad!"
"what happened?"
"I went on stage."
#publicspeaking

Shawna Franc
2012-08-11 14:55:03.0
@Timbecile76
You're weird bestie....I just punched some chick with my new tattoo! She's all bawling screaming, "mommy"....lol

DHannie Cogdell
2012-08-11 17:33:42.0
Wow we get here and it's the wrong fucking theater! REALLY

Stefani Truesdell
2012-08-11 18:49:54.0
Girl time at the movies (@ Megaplex 17 for Spider-Man (2012) w/ 3 others) http://t.co/w7Lh4844

Tommy Ratsinmypants

(a) One week before the soccer game

rio stadium tinto w others
real go rsl sandy ut game
realsaltlake rslvsdal lake
min pic salt team time

http://t.co/8NjnQzGT

Ryan
2012-08-18 16:43:31.0
All hail the king @ Rio Tinto Stadium http://t.co/iMMpTqls

Ryan
2012-08-18 16:47:26.0
Fußball @ Rio Tinto Stadium http://t.co/k7xiOHLZ

Corie Holmes
2012-08-18 17:10:10.0
I am going to the Real Salt Lake game tonight! #superstoked

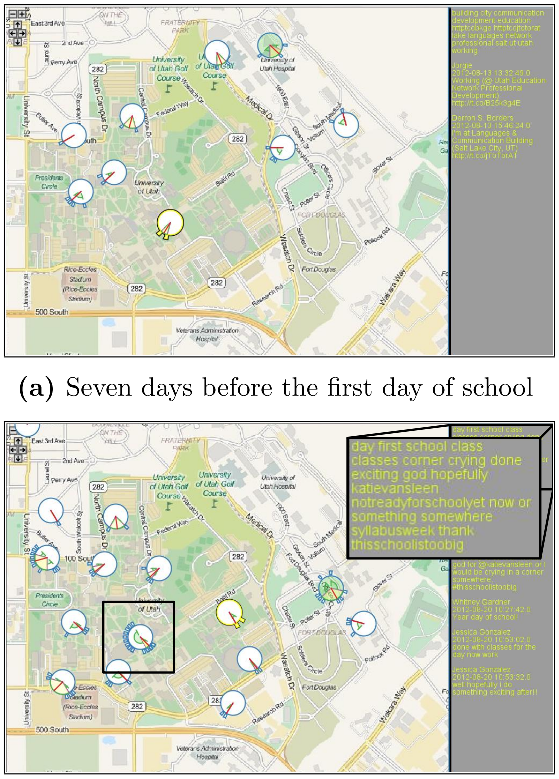
Kyle Reynish
2012-08-18 17:30:21.0
@3dimes well played sir @rodwallace22 @daxmccarty11

Judge Leverich
2012-08-18 17:44:14.0
I'm at Rio Tinto Stadium (Sandy, UT) w/ 7 others http://t.co/BPkAf3BY

?Alexis Tate?
2012-08-18 17:56:30.0
@HannahSKnowles then

(b) The day of the soccer game

Figure 5: Soccer game clyph visualization



building city communication development education httptcobkge httptcojtotorat lake languages network professional salt ut utah working

Jorgie
2012-08-13 13:32:49.0
Working (@ Utah Education Network Professional Development) http://t.co/B25k3g4E

Derron S. Borders
2012-08-13 15:46:24.0
I'm at Languages & Communication Building (Salt Lake City, UT) http://t.co/jToTorAT

(a) Seven days before the first day of school

day first school class classes corner crying done exciting god hopefully katievansleen notreadyforschoolyet now or something somewhere syllabusweek thank thisschoolistoobig

god for @katievansleen or I would be crying in a corner somewhere. #thisschoolistoobig

Whitney Gardner
2012-08-20 10:27:42.0
Year day of school!

Jessica Gonzalez
2012-08-20 10:53:02.0
done with classes for the day now work

Jessica Gonzalez
2012-08-20 10:53:32.0
well hopefully I do something exciting after!!

(b) The first day of school

Figure 6: Visualization of the area surrounding the University of Utah with clyph notches oriented by time

Specifically, we looked for events in areas in and surrounding Salt Lake City for the month of August 2012. For all figures, except Figure 6, the notches in the clyphs are positioned using spatial orientation. In Figure 6, notches are instead oriented by time.

7.1 Short-Run Events

The easiest events to observe with our system are short-term events, so these were the first type of activities we looked for. Several concerts, markets, and sporting events were emphasized by the clyph in our observations. One illustrative event was a Real Salt Lake soccer game which occurred on August 18, 2012. We first noticed an anomaly while moving temporally from the the date a week before the game (Figure 5a) to the date of the game itself (Figure 5b). The visual difference between the clyphs made it apparent that some event was occurring at the location of the stadium.

Upon inspection of the tweets from the clyph centered in this area, we were able to see that almost all of the tweets were focused on events happening at the game. Tight temporal range of the check-ins, coupled with the median mark, enabled us to deduce that the game occurred some time between 17:00 and 22:00, with peak activity at 18:00. Additionally, the close proximity of the quartiles and the range boundaries told us that, as one would expect, nothing is happening at the stadium when it is empty.

7.2 Single Day Events

To identify events on the scale of a single day, we compared clyph placement and composition between proximal days. One of the more prominent events we observed was the first day of school at the University of Utah. We first noticed a spike in activity between the Monday school started and the preceding Friday. Compared to the preceding Monday (Figure 6a), there is an even greater increase in the number of clyphs as well as the per-clyph tweet frequency (Figure 6b). The realization we had with the clyphs for this event was that they were not centralized; that is, after finding a hint of an event from a single clyph in the area, we explored several other clyphs to derive full understanding of the event. This was a logical outcome of the event being more widespread than the soccer game we previously explored. After investigating the text associated with several clyphs, we were able to determine that the event was the first day of school from keywords on the right panel in (Figure 6b). As we further explore the clyphs in Figure 6b, we notice a big spike in the number of tweets in the boxed area over the library. The distribution of clyph notches shows lot of activities happened from 8:00 to 11:00 in the morning and from 13:00 to 22:00 in the evening. The median of the tweets is at around 10:00 on the clyph indicates that there are highly concentrated tweets occurred in the morning. The gap from noon to 13:00 coincides with the fact that the students spread out to have lunch away from library.

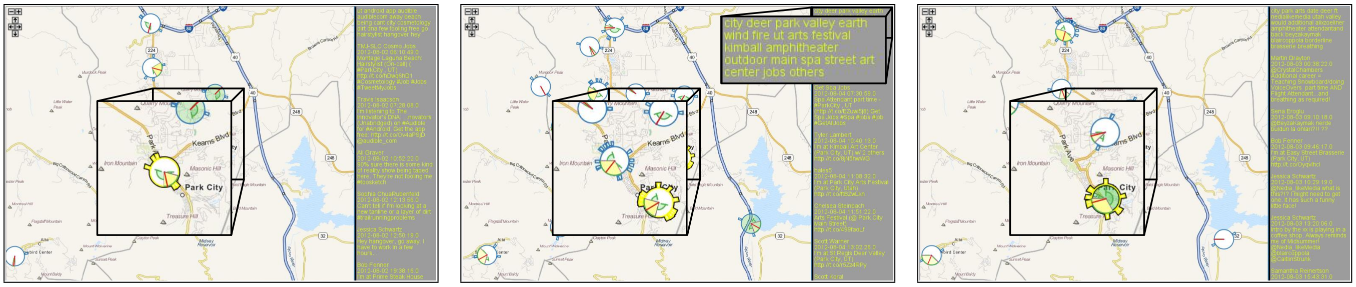
7.3 Multi-Day Events

Festivals were the most common multi-day activity highlighted. The Park City Arts Festival was the first multi-day event we noticed, as there was an observable spike in localized tweet activity during the festival relative to the weekdays surrounding its occurrence (Figure 7). As with the soccer game, we were able to verify that the real time span of the event was well-represented by the median and range data, as the range data encompassed the festival hours within 2 hours, and the median mark occurs roughly in the middle of the festival hours. The quartiles also give interesting insight into when the festival was busiest (approximately 12:00 to 18:00), and they match the intuitive prediction that the timing of this maximum would be in the afternoon hours.

8 Conclusion and discussion

In conclusion, we have presented a novel approach to multiscale visual analysis for geolocated network events. Our approach uses aggregation to support many scales with a novel visual representation, the clyph, to maximize the display of information. We have demonstrated with three case studies that this approach enables users to identify major network events with relative ease.

There remain a few of limitations and associated future work with our approach. The first is that our system still relies upon visual analysis to identify events. Ideally,


city deer park valley earth wind fire ut arts festival kimball amphitheater outdoor main spa street art center jobs others
Park City
Kearns Blvd
Masonic Hill
Iron Mountain
Treasure Hill

(a) The day before art festival **(b)** The first day of festival **(c)** The second day of festival

Figure 7: Clyph visualization of art festival over three days

automatic or semi-automatic machine learning approaches could be leveraged to assist in the visual analysis, reducing the number of visual elements needing inspection. Next, our greedy aggregation algorithm is non-optimal, but worse, it leaves the visual display somewhat sparse. Identifying better approaches to pack glyphs tightly will increase the available information load significantly. The clyph representation also has a few limitations. The clyph assumes a Gaussian distribution in the data. While for a single event, such an assumption seems reasonable, when multiple events occur within a single location, the Gaussian assumption falls apart. More likely a linear combination of Gaussians makes sense; however, other distributions should be investigated as well. Finally, the scale of identifiable events is loosely correlated to the scale of the visualization area, obscuring possibly significant events as the view zooms out. If, for example, we were to look at a visualization of the entire state of Utah, most significant network events would likely appear simply as noise because of the large total number of tweets. Further examination of bottom up analysis methods, which would identify events at a local level and propagate them upward, are necessary.

References

[1] Modest maps. http://www.modestmaps.com/.

[2] U. Brandes and B. Nick. Asymmetric relations in longitudinal social networks. *IEEE Trans. on Vis. and Comp. Graph.*, 17:2283–2290, 2011.

[3] E. Cho, S. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *ACM Know. Dis. and Data Min.*, pages 1082–1090, 2011.

[4] W. Cui, H. Zhou, H. Qu, P. Wong, and X. Li. Geometry-based edge clustering for graph visualization. *IEEE Trans. on Vis. and Comp. Graph.*, 14(6):1277–1284, 2008.

[5] G. Groh, H. Hanstein, and W. Worndl. Interactively visualizing dynamic social networks with dyson. In *Vis. Inter. to the Soc. and the Sem. Web*, 2009.

[6] D. Guo. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Trans. on Vis. and Comp. Graph.*, 15(6):1041–1048, 2009.

[7] D. Guo, J. Chen, A. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Trans. on Vis. and Comp. Graph.*, 12(6):1461–1474, 2006.

[8] D. Guo, M. Gahegan, A. Maceachren, and B. Zhou. Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. *Carto. and Geo. Info. Sci.*, 32:113–132, 2005.

[9] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *IEEE InfoVis*, pages 5–ff, 2005.

[10] D. Holten and J. van Wijk. Force-directed edge bundling for graph visualization. *Comp. Graph. Forum*, 28(3):983–990, 2009.

[11] D. McFarland J. Moody and S. Bender-deMoll. Dynamic network visualization. *American Journal of Sociology*, 110(4):1206–1241, 2005.

[12] D. Keim and H. Kriege. Visualization techniques for mining large databases: A comparison. *IEEE Trans. on Knowledge and Data Engineering*, 8:923–938, 1996.

[13] W. Luo, A. MacEachren, P. Yin, and F. Hardisty. Spatial-social network visualization for exploratory data analysis. In *Workshop on Location-Based Social Networks*, pages 65–68, 2011.

[14] M. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, 2003.

[15] A. Perer and B. Shneiderman. Balancing systematic and flexible exploration of social networks. *IEEE Trans. on Vis. and Comp. Graph.*, 12(5):693–700, 2006.

[16] D. Phan, L. Xiao, R. Yeh, P. Hanrahan, and T. Winograd. Flow map layout. In *IEEE InfoVis*, pages 219–224, 2005.

[17] A. Rae. From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census. *Comp. Environ. and Urban Sys.*, 33(3):161–178, 2009.

[18] C. Reas and B. Fry. Processing.org: a networked context for learning comp. programming. In *ACM SIGGRAPH 2005 Web program*, 2005.

[19] B. Shneiderman and A. Aris. Network visualization by semantic substrates. *IEEE Trans. on Vis. and Comp. Graph.*, 12(5), 2006.