2018

Visualization of Crowd Trajectory, Geospatial Sets, and Audience Prediction at Roskilde Festival 2018

Benjamin Flesch

Ravi Vatrapu

Raghava Mukkamala

Rene Madsen

Visualization of Crowd Trajectory, Geospatial Sets, and Audience Prediction at Roskilde Festival 2018

Case Study for SIGGIS 2018 pre-ICIS Workshop
Location Analytics and Location of Things: Connectedness and Collaboration

Benjamin Flesch, Ravi Vatrapu,
Raghava Rao Mukkamala, René Madsen
Centre for Business Data Analytics (CBDA)
Department for Digitalization
Copenhagen Business School
{bf,rv,rrm,rema}.digi@cbs.dk

Abstract

We present a large-scale study on Geospatial Big Data Analytics in a festival management and crowd safety scenario based on our volunteer work at the largest music festival in Northern Europe, the 2017 and 2018 Roskilde music festival. As large crowds move between concerts, campsites, private parties and public viewing of the FIFA world cup soccer matches across the vast festival area, previously available visualization solutions for the crowd safety staff at Roskilde Festival lack a real-time visualization of crowd trajectory for monitoring of previously known chokepoints and the discovery of potential future chokepoints. We present a real-time visualization of crowd trajectory based on mobile device GPS data that is collected through the festival's smartphone app from a significant subset of all festivalgoers. Hence, we present a recent case for the applications of real-time visualization of geospatial data for large-scale open-air events outside the confinements of urban architecture. Furthermore, we investigate and quantify the phenomenon of festivalgoers staying at the camping areas and sometimes not once visiting the musical performances in the inner perimeters of the festival, thereby finding that up to 3,000 festivalgoers both in 2017 and 2018 chose to not attend musical performances at all. Subsequently, we evaluate a Facebook-based approach on concert audience prediction based on social media audience overlaps between artist and festival Facebook pages, and benchmark its predictive power with a GPS-based measurement of audience sizes based on mobile device GPS data.

Keywords: *Spatial Analysis, Crowd Safety, Real-time Visualization, Crowd Trajectory, Audience Prediction, Crowd Mapping, Geospatial Set Comparison*

Introduction

Large-scale events such as music festivals attract a massive number of visitors, which poses the risk of crowd safety incidents such as accidents, stampedes or even malicious terrorist attacks. In July 2018, our team of data science researchers volunteered to assist Roskilde festival in big data analytics for improvement of festival operations and the generation of actionable insights from the large amounts of data gathered over course of the annual events. As part of this research project, we deployed a Geographic information system (GIS) that analyses visitor GPS data gathered through the festival's own smartphone application in order to improve crowd safety through measures such as real-time crowd monitoring and assessment of crowd trajectory data. Roskilde festival depicts the largest music festival in Northern Europe, having more than 130.000 festival-goers in 2018. During the annual one-week event, a large variety of concerts both from world-renowned artists and upcoming musicians are held. As most of the festival audience stays on-site in one of the camping areas, the festival organizers are concerned with crowd safety around the clock for over a week. The organization behind Roskilde festival depicts an unique composition, as its core staff consists of a small dedicated team of professionals who are augmented by several thousands of volunteers which need to be trained before and continuously managed during the event. In recent years, our team has performed

various data analytics studies at the festival, some of which have been published to the research community (Zimmerman et al. 2016), with many of our findings reported by Danish media (Lykkegaard 2017). Ever since the tragic death of nine festival visitors in a stampede during the Pearl Jam concert at Roskilde music festival in 2000, crowd safety has become an even bigger priority during operations of Roskilde festival in order to prevent any such tragedies from ever occuring again (Vendelo and Rerup 2011). Due to our on-site observation of a crowd density incident at Roskilde festival 2017, which was successfully resolved by the crowd safety operators in due time and without any injuries, we subsequently assessed and enhanced our analysis and reporting capabilities based on the provided GPS data for monitoring of crowd trajectories. This previously undocumented incident concerned crowd movement in opposing directions over a pedenstrian bridge on the festival site, which constituteed the sole link between two major areas of the festival site that are separated by train tracks. Crowd trajectory data analyzed and visualized by our team during the 2018 festival assisted in documentation of similar crowd densitiy events.

Our work contributes to several aspects of crowd monitoring and analytics for modern open-air events such as music festivals. First, our work on the geospatial visualization dashboard which was deployed at the 2018 festival provides crowd safety team with a real-time visualization of crowd trajectory, over the vast expanse of the festival area, and during any weather, lighting and staffing conditions; therefore extending and complementing the current portfolio of GIS tools for availability for the Roskilde Festival crowd safety team. Second, with a total of 45.000 unique devices that provided real-time GPS positions during the 2018 festival period, to our knowledge, this case study contributes the most diverse and exhaustive GPS-based mobility dataset of a large-scale week-long event, with unique attributes of ad-hoc tent cities and pop-up parties contained in the vast open-plan territory at Roskilde festival. Third, the comparative study with the 2017 festival GPS data set showcases unique comparison between different areas of the festival, namely the music area and the campsites. Lastly, we contribute a unique approach on Facebook-based concert audience size prediction and benchmark it with existing GPS-based audience size measurements. This paper is organized as follows. First, we present related work in the fields of geographic information systems (GIS), crowd safety and geospatial movement trajectory visualization. Second, we elaborate on the unique conditions at Roskilde festival 2018 and give further background information on the mass event. Third, we present the data collection methodology and provide insight into the technological foundation of our GIS. Fourth, we present the results of our research with special focus on crowd density heat maps and the real-time crowd trajectory dashboard which was deployed as part of the GIS during the festival. This is followed by a presentation of our findings on the stay-at-the-camp phenomenon which is apparent in data both from 2017 and 2018. Furthermore, we introduce a naive approach on Facebook-based prediction of festival concert audience sizes, and benchmark it against available GPS-based crowd size measurements. Finally, we discuss our results and conclude our work.

Related Work

Research in Geographic information systems (GIS) has further extended into big data and analytics. Enabling technologies such as mobile smartphones have matured and provide real-time global positioning system (GPS) coordinates with spatial patterns that can directly feed into large-scale real-time GIS deployments (Farkas et al. 2016). Analysis and clustering of movement trajectories of individuals has been performed in various city-level case studies (McArdle et al. 2014), with trajectory clustering in open-plan environments mostly focused on naval (Liu et al. 2014) and aerospace applications (Gariel et al. 2011). Geospatial mapping and real-time visualization of large crowds is an evolving topic in research. Crowd density mapping has been performed using bluetooth proximity teachnology that interacts with nearby smartphones (Weppner and Lukowicz 2013). Furthermore, various visual methods for automated congestion detection utilizing video feeds have been proposed (Huang et al. 2015), with implications on crowd safety for large-scale events such as music festivals (Krausz and Bauckhage 2012). Additional research on spatial events of local maxima extracted from social media data such as Flickr photographs and Twitter posts showcases near real-time crowd mapping using non-proprietary data sources as published in Andrienko et al. 2013, pp. 279 ff. Significant research has been performed on geospatial visualization and analytics of cell phone antenna tower GPS location data that is available to telco providers allover the world, where the exact GPS coordinates of the cellphone antenna towers are utilized as proxy for the exact smartphone location. Example trajectory data

studies in Japan analyze anomalous events such as large-scale natural disasters and mass casualty events based on data from 1.5m mobile phones (Witayangkurn et al. 2013). The precision of telco GPS data is limited by the density of cellular antennas, and therefore much less accurate than data from GPS antennas built into modern smartphones. Based on GPS data from mobile phone applications, visualization of crowd density, crowd turbulence, crowd velocity and crowd pressure has been performed in real-time for 800 users (Wirz et al. 2012). Similar small-scale tracking using mobile phone applications has been researched for analysis of tourism flows with ca. 100 devices in South Korea (Yun and Park 2015) and repeat tourist visitor behavior in Hong Kong with ca. 500 devices (McKercher et al. 2012).

Furthermore, the visualization and quantification of crowd movement for large-scale events based on smartphone application GPS data has been studied by several researchers. One of those is the largest event in Switzerland (Blanke et al. 2014), a three-day mass event in Zurich with nearly a million visitors, of which 28.000 have been tracked with GPS. Consequently, work is published on a Dutch dance festival (Daamen et al. 2017) with total of 40.000 visitors, of which a small subset of 109 visitors have been tracked with GPS. Extensive research with focus on recent crowd safety disasters has been published, such as a thorough study of systemic organizational failures that caused the 2010 disaster at Love Parade electronic dance music festival in Germany which left 21 visitors dead (Helbing and Mukerji 2012). To our knowledge, no research on crowd safety aspects of Roskilde music festival has been published in recent years. Work with particular focus on Roskilde festival has been published in 2015 by members of our center, in which a geo-social analysis framework for measurement of stickiness in certain geographic regions of the festival area was proposed (Zimmerman et al. 2016).

Roskilde Festival 2018

Roskilde festival is located in the city of Roskilde in Denmark. It was first established in 1971, and 2018 was its 47th year of activity. The festival area includes a live music area and camping grounds. The total area of the perimeter is 2,500,000 m². Roskilde festival 2018 lasted 8 days between Saturday 30 June and Saturday 7 July 2018. The camping areas opened on Saturday 30 June at 4:00 pm. The inner festival area (including the iconic Orange Stage, Art Zone, Food Court and much more) opened on Wednesday 4 July at 5:00 pm. Headlining artists for the 47th year of Roskilde festival in 2018 were Eminem, Bruno Mars, Gorillaz, Nick Cave & The Bad Seeds, Nine Inch Nails, Massive Attack, David Byrne, Dua Lipe and Nephew. A total of 80,000 full festival tickets and 20,000 one-day tickets were sold for the 2018 festival. The tickets were sold in 69 different countries. The biggest audiences are from Denmark, Norway, Sweden and Germany. 12.5% of the tickets were sold internationally. On top of that, more than 25.000 volunteers received free entry and camping at the festival for several days of volunteer work as part of the festival organization. The festival is located in the small town of Roskilde, and the more than 130,000 festival-goers make Roskilde festival Denmark's fourth largest city measured by population. To our knowledge, no crowd safety incidents have been reported from during the 2018 festival.

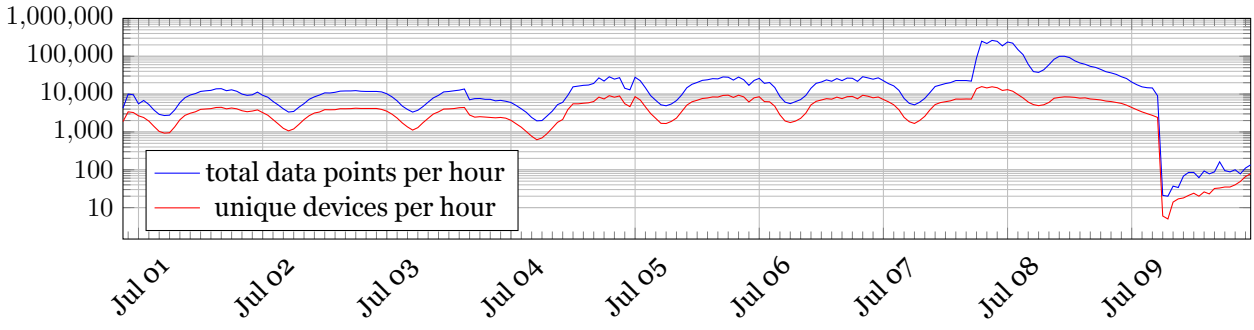


Figure 1. Data collected during 2018 Roskilde festival, sharp drop-off after finish

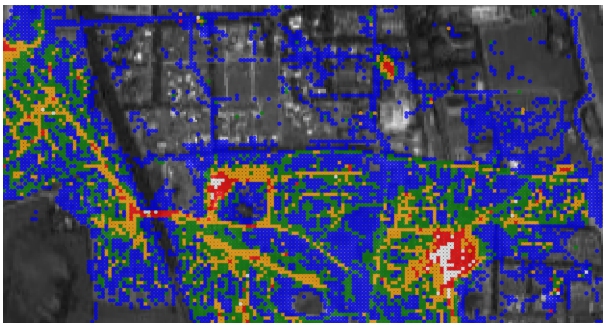
Data Collection

Data was collected in the period from 30th of June to 10th of July 2018 through the festival's official smartphone application based on an opt-in feature of the app. In order to save battery life of the visitors' smartphones, the data was only collected when the Roskilde smartphone application was actively being used. GPS coordinates of the festival area have been demarked with a latitude/longitude rectangle from (55.6656, 12.0273) to (55.5878, 12.1568), only data points within the festival area have been analyzed. The cloud-based data collection system provides a stream of data tuples, consisting of GPS location and timestamp for each device with enabled tracking. As the GPS location is provided by the festival's own smartphone application which runs on the visitors' cell phones, GPS location data is of good quality with little to no scatter. Benefit of fetching visitor GPS data through smartphone applications in comparison with cellular antenna GPS data is that there are no artifacts of bad trajectory data caused by switching the phone connection between neighbouring antennas. (Andrienko et al. 2013, p. 348). Quality of available GPS location data is also positively affected by the fact that the whole festival is held under open sky, so smartphone GPS antennas have little to no visual barriers that could negatively affect satellite signals. During the examination period of the 2018 festival, a total of 45.000 unique devices sent 4.2 million location updates from within the festival area. Figure 1 showcases the number of data points collected during the music festival. The graph displays a repeating pattern with peaks in the late evening and phases of low activity during the early morning hours at around 3 and 4am. The overall peak is located at the final concerts on the evening of July 8th, with more than 120.000 hourly data points collected from 30.000 concurrent devices. The overall low consists of approximately 500 active devices, it is located at 3am on 4th of July, the early morning hours on the day of the opening of the inner festival area which holds the larger concerts. After the festival is concluded with a final concert on July 8th, a Sunday night, a significant, several-magnitude drop-off in activity is observed.

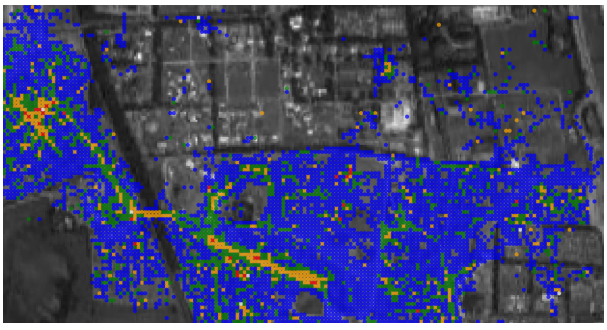
Results

Crowd Density Heatmaps

Crowd density heatmap has been calculated to exemplify two specific events during the festival. Each heatmap depicts a map of the main festival area including the major campsites for festival visitors. Campsites are located to the west, south and south east of the main festival area. Each pixel of the heatmap depicts a 10x10 meter grid cell. For each grid cell, the aggregate number of seconds that visitors spent in that cell during a certain time frame is calculated. A color is chosen based on the aggregate number of seconds spent within each cell. Colors range from transparent color for zero-second cells, via green, yellow, red to white color depicting the maxima.



(a) During world cup soccer match Denmark vs. Croatia on July 1st 2018 at 19:30-23:00



(b) After finish of the soccer match with a win for Croatia at 23:00-2:00 local time

Figure 2. Heat maps of festival crowd density during and after Denmark vs. Croatia match

Crowd Density Heatmap of World Cup Soccer Match Denmark vs. Croatia

Figure 2 showcases crowd density heatmap for the Denmark versus Croatia world cup soccer match on 1st of July 2018 at 20:00 local time. As soccer fans might know, Denmark lost the nearly three-hour match after a fierce penalty shootout with 3:2 goals. Nevertheless, thousands of Danes watched the live broadcast of the soccer match in a public viewing area on Roskilde festival. It must be noted, that at this point in time the main festival area is still closed to the public, therefore the upper half of the heatmap seems only very sparsely populated.

Figure 2a depicts the crowd density heatmap during the soccer match in the 19:30 to 23:00 time frame. The heatmap communicates the sheer size of the public viewing audience, which is located in the campsite area at the bottom right of the visualization. The intensity of the gathered crowd is illuminated in white and red colors. Furthermore, Countdown stage is displayed as a white peak at the center of the heatmap. Left of Countdown stage crowd a high crowd density is visible at the bridge crossing the train tracks that separates the two major areas of Roskilde festival. At the very center top of the heatmap, a red-color density peak depicts the backstage and volunteer area, where a lot of volunteers gather to watch the soccer game on non-public television screens. Major pathways through the campsites are apparent in the heat map.

Figure 2b depicts the crowd density heatmap after the soccer match is concluded with a loss for Denmark. The exact timeframe of this heatmap is 23:00 until 2:00 the next day. Crowd density at the site of the public viewing of the soccer match has dropped back to normal levels. The major east-west pathway leading to the bridge that crosses the train tracks still exhibits major activity. This activity is mainly caused by open-air music playing from private speaker boxes in that area. At the top left corner of the heatmap, further campsite party activity shows high crowd density well into the early morning hours. Activity in the backstage and volunteer area at the center top of the heatmap has subsided. The concert at Countdown stage is also finished, therefore the crowd has dispersed. Both heatmaps display a significant shift of crowd density accross the festival area. The pedestrian bridge, major choke point in the east-west connection, consistently shows global maxima in crowd density. The lack of information on crowd trajectory depicts a limitation of the crowd density heatmaps. The overall direction of crowd movement is not visible in the visualizations, therefore no predictions on future maxima in crowd density can be made given these figures.

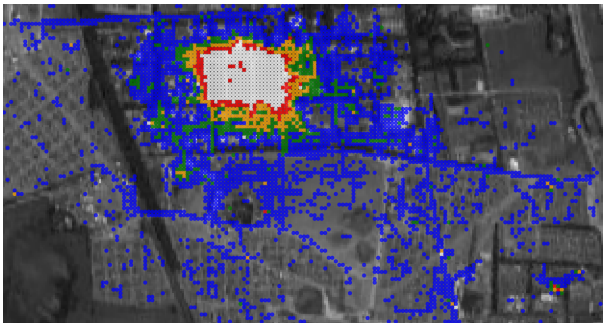
Crowd Density Heatmap of EMINEM Concert

Figure 3 showcases crowd density heatmaps for the concert of EMINEM, main headliner of Roskilde festival 2018, at the Orange stage on 4th of July. Festival administration assumes that around 100.000 visitors attended the concert which started at 22:00.

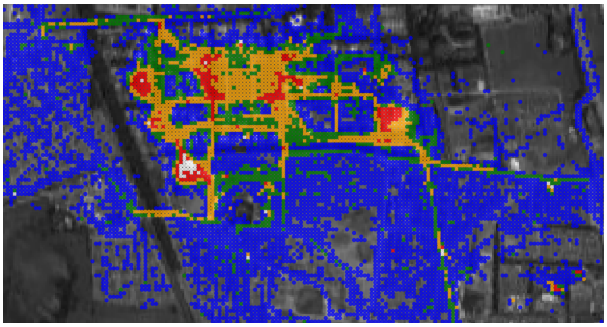
Figure 3a depicts the crowd density heatmap during the concert at Orange stage, the major stage of the festival. It is striking that a large area of the festival site shows the highest crowd density in white color, this is where Orange stage is located. The campsites which were thoroughly illuminated in the previous example of figure 2 show only very sparse crowd density. When observing this heatmap, we can conclude that the large majority of festival visitors is attending the EMINEM concert.

Figure 3b depicts the crowd density heatmap in the hour immediately after the conclusion of the EMINEM concert at around 23:00. The major peak in crowd density has dispersed and several new peaks have formed at the various stages in the core festival area. Apollo stage is the only stage that showcases a significant white-color peak in crowd density at this point of time. Arena stage at the center right of the heatmap showcases a second, albeit not as intense peak in crowd density. The major passageways and egress routes towards the campsites show <activity. No major peaks within the campsites become apparent in the heatmap. When comparing this heatmap with previous figure 3a we observe that the densely crowded audience of the EMINEM concert has dispersed from Orange stage and overall crowd density has averaged out in the core festival area.

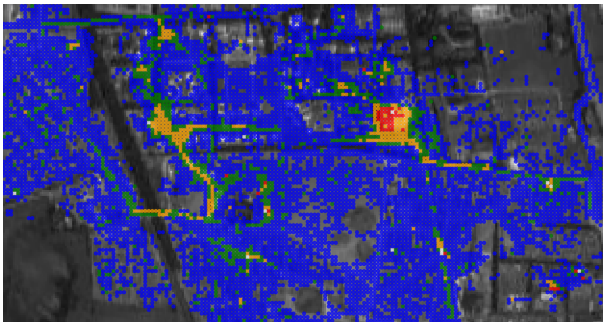
Figure 3c depicts the crowd density heatmap between midnight and 1am. The crowd has further dispersed and most have returned to the campsites. A significant crowd density is to be observed at Arena stage at the center right of the heatmap, where one of the last concerts for this day is being performed. Major passageways show high levels of crowd density, and some areas of the campsites show crowd densities around



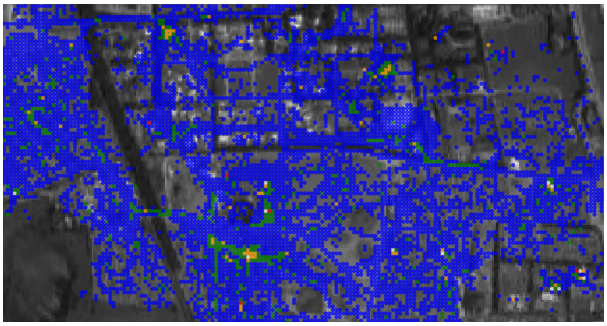
(a) July 4th from 22:00-23:00 local time



(b) July 4th from 23:00-24:00 local time



(c) July 4th from 24:00-1:00 local time



(d) July 5th from 1:00-2:00 local time

Figure 3. Crowd density heatmaps during and after EMINEM concert

major party hostspots.

Figure 3d depicts the crowd density heatmap between 1am and 2am. The last concert has finished, and no crowd density peak is observed within the core festival area. Minor peaks are visible at the party hotspots in the campsite. There is low activity on the pedestrian bridge crossing. Most vistors have retreated to their tents. The four one-hour crowd density heatmaps during and after the EMINEM concert, one of the largest concerts at Roskilde festival 2018, clearly communicate crowd density peaks during and after the concert. Furthermore, figure 3b illuminates major egress routes from the main festival area towards the campsites directly after the concert concluded. Due to clever choice of observation time frames we can visualize the shift in crowd density across the festival area. Nevertheless, information on individual instances of crowd trajectory are not apparent from the visualization. A real-time observation or prediction of shifting crowd density is not possible with the heatmaps at hand.

Visualization of Real-Time Crowd Trajectory Data

Due to the limitations of the crowd density heatmaps presented in the previous section, a visualization of real-time crowd trajectory is used to display and predict crowd movements based on the incoming GPS points from users of the mobile app. For this purpose, a web-based visualization dashboard was provided for the festival crowd safety team, as showcased in this section. The geospatial visualization at hand utilizes the two most recent data points for each device, thereby creating a movement vector to display each individual trajectory. In order to distinguish between compass direction of the movement, a compass dial (or wind rose) was used for color-coding the directions in which the crowd is heading. The opacity of the trajectory vector is determined by the age of the underlying data points, providing means to visually distinguish between the most recent and past data points. The visualization is updated every five seconds, therefore the most recent movement data is always displayed. Crowd trajectories become apparent through the visual overlapping of several dozen or hundred movement vectors. A limit on the maximum length of individual movement vectors shows more granular features of the crowd movement and can be utilized when zooming in to the map. This crowd trajectory visualization allows the viewer to visually distinguish between real-time crowd

Figure 4. Visualization of real-time crowd trajectory and movement of visitors between festival site and city center of Roskilde in our web-based dashboard.

movements in different directions based on color, therefore enabling the early discovery of potential future chokepoints that are not visible in more traditional heatmaps. Based on the high-quality GPS location data from smartphone devices and the good coverage of nearly 50% of all visitors, an effective visualization is achieved that can be operationally used by crowd safety experts.

Figure 4 depicts real-time crowd trajectory on the festival area and the surrounding campsites. Major pathways are clearly visible, and the pedestrian bridge is identifyable as a chokepoint where crowd movements in opposing directions meet. Shifts in crowd density can be observed for example at the ingress route from the north entry to the main festival area. A second shift in crowd density can be observed in the purple westward crowd trajectory at the bottom right part of the visualization, where the major east-side entrance to the campsites is located. Furthermore, figure 4 depicts crowd trajectory of the festival site and surrounding parts of the city of Roskilde including the adjacent highway. The visualization depicts the trajectory of festival visitors who travel between Roskilde festival and the city center of Roskilde situated at the very top of the image. The blue color of the trajectory vectors in the streets following up to the city center indicate overall crowd trajectory in those areas is headed towards north, where the city center is located. At the south-west to north-east leading highway, clear trajectory data is shown from both lanes, with movement vectors of cars heading north-east in greenish color, and cars heading south-west in reddish color. The pedestrian bridge across the railway tracks in the festival area is clearly visible, the purple to blue color of the surrounding crowd trajectory vectors indicates that there is significant crowd movement in north-west direction across the bridge. No significant eastward movement across the pedestrian bridge can be observed.

Geospatial Set Analysis: Camping vs. Music Area

In recent years, our festival analytics team has observed through interviews with festivalgoers that several visitors do not plan to attend concerts at any given day, and that some visitors even remain at the campsites

Areas	2017	2018
Total Festival Area	**51,109**	**37,755**
Music Area	**44,453**	**33,871**
% share of total	*87%*	*90%*
# missing of total	*6,656*	*3,884*
Camping Area	**48,802**	**36,634**
% share of total	*95%*	*97%*
# missing of total	*2,307*	*1,121*

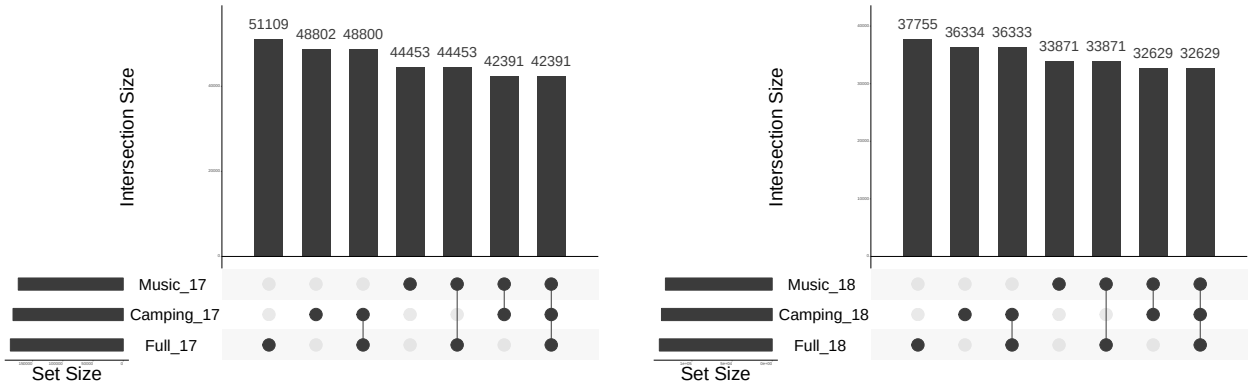
Figure 5. Comparison of unique devices in geospatial sets between camping and music areas.

throughout the whole festival, hence visiting the music area in the inner perimeter of Roskilde festival not even once. In order to quantify this effect, a geospatial set analysis has been performed. For this purpose, we can divide Roskilde festival into two major areas, access to both is included with purchase of the festival entrance ticket. One is the camping area, which is open throughout the festival week, and the other is the music area, which is the inner perimeter and only accessible for visitors from Wednesday to Sunday. Based on geofences for both areas, we can quantify the number of unique users who attended the actual musical performances in the music area of the festival site, and compare it with the total number of unique users that were observed in the camping area.

Figure 5 showcases the results of this comparison for the years 2017 and 2018. We observe that in 2018, the number of unique users from which GPS data has been collected is signifiantly less than in the previous year. Furthermore, it becomes apparent that in both years the number of unique users observed in the camping area is higher by around 3,000 than the number of unique users observed in the music area. This finding seems to underline the above-mentioned effect that a significant number of festivalgoers (or their devices) never make it to the main music area during festival week.

Figure 6 showcases the geospatial set intersections of unique GPS devices between the full festival perimeter, the music area, and the camping area for both years 2017 and 2018. In 2017, a total of $42,391$ unique devices have been both in camping and music area, which means from the $48,802$ devices that have been observed in the camping area, a total of $48,802 - 42,391 = 6,411$ devices have never made it from the campsites to the music area. In 2018, a total of $32,629$ unique devices have been in both areas, a total of $36,334$ only in camping area, and therefore a total of $36,334 - 32,629 = 3,705$ devices have never made it from the campsites to the music area.

Figure 7 underlines this observation. For any given day of the festival, a significant part of the overall GPS devices does not touch the music area at all. In 2017 it is between 12 and 14% of all devices, and in 2018



(a) Set Intersections in 2017 **(b) Set Intersections in 2018**

Figure 6. Upset-style visualization of geospatial set intersections at Roskilde Festival.

		Camping Area Open			Camping Area & Music Area Open				
Year	Areas	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
2018	Total Festival Area	13,868	14,140	15,985	24,889	28,684	27,746	24,633	11,931
	Music Area	3,051	3,199	4,098	20,310	25,139	23,632	20,587	9,030
	% of total	*22%*	*23%*	*26%*	*82%*	*88%*	*85%*	*84%*	*76%*
	Camping Area	13,483	13,779	15,509	21,962	25,744	24,992	22,064	10,051
	% of total	*97%*	*97%*	*97%*	*88%*	*90%*	*90%*	*90%*	*84%*
2017	Total Festival Area	17,192	17,993	21,013	31,806	34,062	33,077	28,479	9,978
	Music Area	2,108	2,389	3,353	27,309	29,193	28,387	25,038	6,147
	% of total	*12%*	*13%*	*16%*	*86%*	*86%*	*86%*	*88%*	*62%*
	Camping Area	16,773	17,553	20,363	28,952	31,024	29,414	24,745	7,963
	% of total	*98%*	*98%*	*97%*	*91%*	*91%*	*89%*	*87%*	*80%*

Figure 7. Daily comparison between geospatial sets across camping and music areas.

between 12 and 18% of all unique devices. In both years, attendance of the music area peaks on Thursdays with 23,632 in 2018, down from 28,387 in 2017.

Audience Prediction of Festival Concerts based on Facebook Data

We attempt to predict the audience for each artist playing at the festival based on a naive application of Social Set Analysis methodology Vatrapu et al. 2016 on Facebook activity. First, we fetch all public data from the Roskilde Festival Facebook page for the six months leading up to the festival. Then, we fetch all public data in the same time period from the Facebook pages for all artists that are scheduled to play at Roskilde Festival. Next, a pairwise set intersection is performed between users of the Roskilde Festival Facebook page and each artist's Facebook page. This way, a metric is generated that reflects the overlap of social media audience between the Facebook pages of Roskilde Festival and each of the artists. The cardinality of each set intersection is documented and sorted, resulting in a list of the top 15 largest audience overlaps between the festival and its artists. The underlying assumption is that Facebook activity is a good proxy for real-world interactions for both the festival and the artists, therefore, if users are active both on the artist

Prediction w/ Facebook Data			Error in Predicted Rank	Measurement w/ GPS Data		
Artist	Rank	# unique		Rank	# unique	Artist
Foo Fighters	1	3,269	+ 1	1	15,834	Father John Misty
Arcade Fire	2	2,053	+ 6	2	13,081	Foo Fighters
Phlake	3	2,046	+ 2	3	11,132	The Weeknd
Karl William	4	1,738	+ 6	4	9,948	The XX
Emil Stabil	5	1,734	+ 6	5	9,814	Phlake
The Hellacopters	6	1,719	+ 7	6	9,540	Royal Blood
Sort Sol	7	1,574	+ 8	7	8,240	Trentemöller
The Weeknd	8	1,349	- 5	8	7,074	Arcade Fire
Royal Blood	9	1,226	- 3	9	6,332	Carl Emil Petersen
Anthrax	10	1,103	+ 2	10	4,097	Karl William
Justice	11	1,076	+ 3	11	3,121	Emil Stabil
Father John Misty	12	1,050	- 11	12	1,623	Anthrax
Trentemöller	13	956	- 6	13	1,066	The Hellacopters
The XX	14	946	- 10	14	665	Justice
Carl Emil Petersen	15	919	- 6	15	196	Sort Sol

Figure 8. Prediction of TOP15 artists at Roskilde Festival 2017 based on pairwise audience overlaps between festival and artist Facebook pages in line with Social Set Analysis methodology, contrasted with GPS-based audience size measurements for each of the artists.

and the festival pages, we should be able to quantify this shared audience, and be able to rank attractiveness of each artist for the festival audience based on this metric.

Figure 8 showcases our prediction of the Top 15 artists of Roskilde Festival 2017 based on above-mentioned naive approach utilizing Facebook audience overlaps. This prediction is benchmarked against a geospatial measurement of audience size for each artist based on the number of unique GPS devices present at the particular stages during the time period of each artist's concert. Please note that due to limited availability of Facebook data, this prediction has only been made for the 2017 festival. It becomes apparent that the overall predictive strength of this naive approach is very limited. Even though we observe that the predicted rank within the Top 15 artists is most correct for Foo Fighters, Phlake, Anthrax and Justice, three of the top four largest audiences, namely Father John Misty, The Weeknd and The XX, were severely underestimated by the Facebook-based prediction. Based on the GPS-based measurement of audience sizes, Father John Misty had the overall largest concert at Roskilde Festival 2017.

Discussion and Conclusion

Crowd Heatmaps and Crowd Trajectory Visualizations

During the course of our multi-year research project with Roskilde festival, we have identified several limitations of the utilization of heatmaps for crowd safety purposes in a large-scale event, in particular that the real-time or near-real-time monitoring of potential choke points is not possible using the off-the-shelf GIS that is available to the festival crowd safety operators. This limitation is underlined by the fact that the area contains a major chokepoint, namely the pedestrian bridge across the train tracks diving the festival area in two, that arises from the naturally given terrain of the festival perimeter. To address the need for a real-time visualization of crowd trajectories in known chokepoints and the visual detection of previously unanticipated chokepoints with special regard to the specific circumstances of Roskilde Festival, we proposed and implemented a web-based dashboard using real-time GPS data points which enables crowd safety operators at Roskilde Festival to track high-level crowd movement and potential chokepoints in order to potentially preempt and analyze critical situations. Deployment of this real-time GIS based on smartphone GPS data feeds was performed by a small team of researchers using off-the-shelf equipment. The web-based visualization dashboards provides scalable and accessible data for other members of the festival organization. Therefore, and based on the above, this work provides a novel contribution to mobile-based location analytics and spatial decision support for large-scale events such as music festivals.

Geospatial Set Analysis: Camping vs. Music Areas

Through a comparison of geospatial sets, we were able to investigate the self-reported phenomenon that not all festivalgoers actually visit the concerts in the music area in the inner perimeter of the Roskilde Festival. Our findings suggest that in both 2017 and 2018, a significant number of GPS devices have never visited the music area at all, therefore underlining the first empirical observation of such a behavior. Furthermore, it was shown that in both years and during the whole festival period with musical performances, the number of unique data points in the camping areas was always larger than the number of unique data points in the music area, with the exception of a single day, Saturday 2017. Due to space restrictions, this study is unable to go into greater detail concerning the underlying reasons for this observation, and further qualitative studies are needed to evaluate the reasons behind this behavior. Based on our multi-year experience with this particular festival we suggest two main reasons behind this phenomenon. First, a number of festivalgoers really doesn't like the musical entertainment that is provided in the music area and prefers the parties at the camping area, as it was evidenced throughout several in-person interviews. Second, it must be noted that a potentially significant number of festivalgoers might leave their mobile phones behind at the campsites when attending any of the concerts in the music areas, a behavior that could be further quantified through adequate interviews in the future.

Prediction of Concert Attendance based on Facebook data

Using Social Set Analysis methodology, the 2017 Roskilde Festival concert attendance was predicted based on overlapping of social media audience of each artist's Facebook page with the Roskilde Festival page. Our findings suggest that the results of such a naive approach on prediction of festival concert audience sizes does not hold when benchmarked against GPS device measurement data, which in our case acts as ground truth data on relative audience sizes. We find that with some artists such as Foo Fighters, Phlake and Anthrax, the Facebook-based prediction gives a very good indication of overall audience size at the festival. On the other hand, Father John Misty gathered the largest measured audience of 15,834 festivalgoers, en par with Foo Fighters' 13,081 audience members, but the Facebook-based prediction severely underestimated this and other artists, only measuring one third of the social media audience of Foo Figthers in Facebook. Therefore, further research is needed to address the impact of non-included variables into the predictive model. Such variables could be the amount of promotional activities in social media by Roskilde festival for each artist, the overall festival day and time slots of each artist's concert, and cannibalization effects between artists.

Acknowledgements

This work results from our big data analytics collaboration with Roskilde festival. We thank the Roskilde festival organizers, the local authorities involved in crowd safety measures at the festival site and various volunteers from both our research group and the festival team.

References

Andrienko, G., N. Andrienko, P. Bak, D. Keim, and S. Wrobel (2013). *Visual Analytics of Movement*. eng. Berlin, Heidelberg: Springer Berlin Heidelberg Imprint: Springer.

Blanke, U., G. Troster, T. Franke, and P. Lukowicz (2014). "Capturing crowd dynamics at large scale events using participatory gps-localization". In: *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2014 IEEE Ninth International Conference on*. IEEE, pp. 1–7.

Daamen, W., E. Kinkel, D. Duives, and S. P. Hoogendoorn (2017). *Monitoring Visitor Flow and Behavior During a Festival: the Mysteryland Case Study*. Tech. rep.

Farkas, D., B. Hilton, J. Pick, H. Ramakrishna, A. Sarkar, and N. Shin (2016). "A Tutorial on Geographic Information Systems: A Ten-year Update." In: *Communications of the Association for Information Systems* 38.1.

Gariel, M., A. N. Srivastava, and E. Feron (2011). "Trajectory clustering and an application to airspace monitoring". In: *IEEE Transactions on Intelligent Transportation Systems* 12.4, pp. 1511–1524.

Helbing, D. and P. Mukerji (2012). "Crowd disasters as systemic failures: analysis of the Love Parade disaster". In: *EPJ Data Science* 1.1, p. 7.

Huang, L., T. Chen, Y. Wang, and H. Yuan (2015). "Congestion detection of pedestrians using the velocity entropy: A case study of Love Parade 2010 disaster". In: *Physica A: Statistical Mechanics and its Applications* 440, pp. 200–209.

Krausz, B. and C. Bauckhage (2012). "Loveparade 2010: Automatic video analysis of a crowd disaster". In: *Computer Vision and Image Understanding* 116.3, pp. 307–319.

Liu, B., E. N. de Souza, S. Matwin, and M. Sydow (2014). "Knowledge-based clustering of ship trajectories using density-based approach". In: *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, pp. 603–608.

Lykkegaard, A. M. (2017). "Researchers know how you behave at Roskilde Festival". In: *CBS Wire*.

McArdle, G., U. Demšar, S. van der Spek, and S. McLoone (2014). "Classifying pedestrian movement behaviour from GPS trajectories using visualization and clustering". In: *Annals of GIS* 20.2, pp. 85–98.

McKercher, B., N. Shoval, E. Ng, and A. Birenboim (2012). "First and Repeat Visitor Behaviour: GPS Tracking and GIS Analysis in Hong Kong". In: *Tourism Geographies* 14.1, pp. 147–161. eprint: https://doi.org/10.1080/14616688.2011.598542.

Vatrapu, R., R. R. Mukkamala, A. Hussain, and B. Flesch (2016). "Social Set Analysis: A Set Theoretical Approach to Big Data Analytics". In: *IEEE Access* 4, pp. 2542–2571.

Vendelo, M. and C. Rerup (2011). "Sensegiving and crowd safety after the pearl jam concert accident". In: *The Academy of Management Annual Meeting, San Antonio, TX*.

Weppner, J. and P. Lukowicz (2013). "Bluetooth based collaborative crowd density estimation with mobile phones". In: *Pervasive computing and communications (PerCom), 2013 IEEE international conference on*. IEEE, pp. 193–200.

Wirz, M., T. Franke, D. Roggen, E. Mitleton-Kelly, P. Lukowicz, and G. Tröster (2012). "Inferring crowd conditions from pedestrians' location traces for real-time crowd monitoring during city-scale mass gatherings". In: *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2012 IEEE 21st International Workshop on*. IEEE, pp. 367–372.

Witayangkurn, A., T. Horanont, Y. Sekimoto, and R. Shibasaki (2013). "Anomalous Event Detection on Large-scale GPS Data from Mobile Phones Using Hidden Markov Model and Cloud Platform". In: *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*. UbiComp '13 Adjunct. Zurich, Switzerland: ACM, pp. 1219–1228.

Yun, H. J. and M. H. Park (2015). "Time–space movement of festival visitors in rural areas using a smart phone application". In: *Asia Pacific Journal of Tourism Research* 20.11, pp. 1246–1265.

Zimmerman, C., R. Madsen, H. H. Eliassen, and R. Vatrapu (2016). "Space vs. Place: Comparing Space-based Movements and Place-based Experiences at the Roskilde Festival 2015". In: *Proceedings of the 7th 2016 International Conference on Social Media & Society*. SMSociety '16. London, United Kingdom: ACM, 11:1–11:10.