Computers, Environment and Urban Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Linking cyber and physical spaces through community detection and clustering in social media feeds

Arie Croitoru [a,*], N. Wayant [b], A. Crooks [c], J. Radzikowski [a], A. Stefanidis [a]

[a] The Center for Geospatial Intelligence, Dept. of Geography and Geoinformation Science, George Mason University, 4400 University Drive, MS 6C3, Fairfax, VA 22030, United States
[b] US Army Geospatial Research Laboratory, 7701 Telegraph Road, Alexandria, VA 22315-3802, United States
[c] Dept. of Computational Social Science, George Mason University, 4400 University Drive, MS 6C3, Fairfax, VA 22030, United States

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Social media
Spatiotemporal clustering
Social network analysis
Community detection
Geospatial analysis

ABSTRACT

Over the last decade we have witnessed a significant growth in the use of social media. Interactions within their context lead to the establishment of groups that function at the intersection of the physical and cyber spaces, and as such represent hybrid communities. Gaining a better understanding of how information flows in these hybrid communities is a substantial scientific challenge with significant implications on our ability to better harness crowd-contributed content. This paper addresses this challenge by studying how information propagates and evolves over time at the intersection of the physical and cyber spaces. By analyzing the spatial footprint, social network structure, and content in *both* physical and cyber spaces we advance our understanding of the information propagation mechanisms in social media. The utility of this approach is demonstrated in two real-world case studies, the first reflecting a planned event (the Occupy Wall Street – OWS – movement's Day of Action in November 2011), and the second reflecting an unexpected disaster (the Boston Marathon bombing in April 2013). Our findings highlight the intricate nature of the propagation and evolution of information both within and across cyber and physical spaces, as well as the role of hybrid networks in the exchange of information between these spaces.

1. Introduction

The past few years have witnessed the dramatic increase in the adoption and use of social media (Kaplan & Haenlein, 2010). In the U.S. alone, approximately two-thirds of online users participate in social media (Smith, 2011), spending on average between 3.6 and 6.5 h a month in social networking sites such as Facebook or Twitter (Nielsen, 2012). This has led to an unprecedented increase in the volume of data generated by social media users: every minute we have over 270,000 tweets (or retweets) contributed worldwide (Forbes, 2012), 3000 images posted in Flickr (Sapiro, 2011), and 100 h of video uploaded in YouTube (YouTube, 2014). These are but a few examples of the shift that has occurred in recent years toward user-generated digital content. With millions of users around the world, this trend is likely to further intensify (Hollis, 2011) as technological advances empower users to contribute richer data at higher rates.

Social media services and platforms offer a wide array of digital channels for expression and interaction, ranging from forums/message boards (e.g. MacRumors), weblogs (e.g. Blogger, Wordpress), and microblogging (e.g. Twitter, Tumblr, Weibo), to wikis (e.g. Wikipedia, Wikimapia), social networking services (e.g. Facebook, Google+, LinkedIn), and podcasts (Video and Audio e.g. iTunes, Ustream). Such media have enabled the general public to contribute, disseminate, and exchange information (Kaplan & Haenlein, 2010), by introducing a bottom-up alternative to complement the traditional top-down nature of Web 1.0 (Schneckenberg, 2009). This has not only resulted in a change in traditional journalism and news reporting (Deuze, 2008; Kwak, Lee, Park, & Moon, 2010), but it is also leading to new opportunities within the geographical sciences (Caverlee, Cheng, Sui, & Kamath, 2013; Sui & Goodchild, 2011) due to the rich geographic context and context social media data often provides. A noteworthy example of this trend is the livehoods project (Cranshaw, Schwartz, Hong, & Sadeh, 2012) that is used to characterize and understand urban dynamics using social media. Indeed, social media, and micro-bogging in particular, have already been shown useful in predicting pandemics (Chunara, Andrews, & Brownstein, 2012; Culotta, 2010; Ritterman, Osborne, & Klein, 2009) or natural disasters

* Corresponding author.
E-mail addresses: acroitor@gmu.edu (A. Croitoru), Nicole.M.Wayant@usace.army.mil (N. Wayant), acrooks2@gmu.edu (A. Crooks), jradziko@gmu.edu (J. Radzikowski), astefani@gmu.edu (A. Stefanidis).

http://dx.doi.org/10.1016/j.compenvurbsys.2014.11.002
0198-9715/

(e.g. Corbane, Lemoine, & Kauffmann, 2012; Crooks, Croitoru, Stefanidis, & Radzikowski, 2013; Zook, Graham, Shelton, & Gorman, 2010) to name a few.

As we increasingly embrace the use of crowd-contributed content, gaining a better understanding of how physical space events are reported and discussed within these hybrid communities is a substantial theoretical challenge that also has significant application potential. This paper addresses this challenge by studying how information propagates and evolves over time at the intersection of the physical and cyber spaces, considering representative test cases and studying them under the *lens* of geosocial analysis. By analyzing the spatial footprint, social network structure, and content in *both* physical and cyber spaces we can advance our understanding of the complex mechanism through which information regarding localized events is propagated through social media.

A particularity of social media that renders such study necessary is the fact that, unlike other forms of volunteered geographic information, contributions there are part of a networking process, whereby individuals share and exchange information with other members of these online communities (Stefanidis et al., 2013). This networking activity may center around a variety of topics, ranging from personal observations on minutia to commentaries on issues of broader interest (Aiello et al., 2013; Mischaud, 2007). Understanding how people participate in this process remains a substantial, cross-disciplinary theoretical challenge. As a way to address this issue, Farnham and Churchill (2011), for example, discussed the issue of *cyber (online) presences*, governed by principles of cyber interaction and information flow. However, as these studies emerged from the social psychology domain, they often fail to adequately address the role of the physical space in these cyber interactions. People still live and function primarily in a physical space (rather than the cyber one), and their interactions in this space still play a central role in shaping their behavior. As social media becomes an integral part of our societies, understanding the interplay between cyber presence and the corresponding physical space (so called "polysocial reality" (Applin & Fischer, 2012)) becomes increasingly important, as it will elevate our capability to leverage such content for a variety of purposes.

Mapping and understanding the relations between the cyber and the physical spaces, and in particular the information flow between them is a substantial scientific research challenge, as these two spaces are often not explicitly related, nor are they studied in tandem. The emergence of geolocated social media presents a unique opportunity to address this challenge by allowing us to link cyber and physical activities through user interactions, and understand how peoples' actions and reactions to events manifest themselves across these spaces. Such knowledge is critical in a wide range of applications of broader societal value (e.g. disaster response), providing additional motivation for this research.

Our focus is on studying the connections between the cyber and the physical spaces (especially as it relates to reports of events in the physical space), as they are expressed through social associations *and* physical proximity. By doing so we will show how we can identify connections *across* these two spaces, and demonstrate the value of studying both simultaneously rather than separately. We argue that by studying social networks in *both* physical and cyber spaces, combining social network analysis (SNA) and spatiotemporal data clustering we can better understand their complex structure. While SNA is a rapidly growing field (e.g., Newman (2010) and Barabási (2012)), it is only recently emerging as a tool in geospatial analysis, and is often underutilized (Ter Wal & Boschma, 2009). Moreover, SNA too is weakened because of the lack of geographic consideration when exploring social relations (e.g. Bosco, 2006). Only recently have we started seeing some early studies that attempt to infuse geography into this issue, addressing for example the geographic scope of topics discussed in on-line communities (HerdaGdelen, Zuo, Gard-Murray, and Bar-Yam (2013)). Our work contributes to this issue by linking the cyber and physical spaces through SNA and spatiotemporal analysis, aiming to bridge the gap between these two fields.

The remainder of this paper is organized as follows. In Section 2 we discuss the rise of geosocial media as a new social communication avenue and a novel source of geosocial information. In particular, we discuss the notion of physical presence within social media and its importance for exploring the relation between the cyber and the physical domains. In Section 3 we discuss how communities and groups can be detected in both the cyber and physical space, and how they can be processed to form a 'hybrid' geosocial view of communities. To showcase these concepts and their benefits, in Section 4 we present the analysis of two case studies that make use of Twitter data associated with two different types of events: a planned activity during the Occupy Wall Street (OWS) Day of Action (November 17th, 2011), and the response to the Boston Marathon Bombing (April 15, 2013). This paper is concluded with a summary and outlook in Section 5.

2. The rise of geosocial media and spatial presence

The power of social media to disseminate information of societal importance has been showcased over the last few years with respect to a range events, from citizen journalism (e.g. the 2008 Mumbai terrorist attacks; Arthur, 2008), to civil unrest (e.g. the 2011 London riots (Glasgow, Ebaugh, & Fink, 2012) and the 'Arab Spring' (Christensen, 2011; Howard, Agarwal, & Mazammil, 2011)), military operations (e.g. the 2011 U.S. raid on Bin Laden's hideaway (McCullagh, 2011)) and health (e.g. Culotta, 2010). Within this wide range of themes social media content often relates to the real-world space and the events occurring in it, it may contain some geolocation information, thus making it *geosocial media*. The emergence of such information represents a complementary approach to what was initially considered to be Volunteered Geographic Information (VGI; Goodchild, 2007; Sui, 2008), as it focuses on geographic information that may be conveyed through social media but not intentionally volunteered.

Generally, geolocation information in social media may be available at two different levels of granularity: in the form of precise coordinates (e.g. GPS), or through a toponym detected in the media content (e.g. mentioning a city name) that can be resolved using a gazetteer (e.g. Croitoru, Crooks, Radzikowski, & Stefanidis, 2013; Fink et al., 2009). This information is typically harvested together with the rest of the content from the corresponding social media platform Application Programming Interfaces (APIs). Alternatively, geolocation information can be harvested from the content of the users' profiles (e.g. establishing a user's home location), however this information tends to be less reliable as it does not necessarily reflect user location at the moment that a particular contribution was made (and is therefore not used in our study). It is worth noting that reports on the percentage of geolocated social media – and twitter in particular – vary. For example, Java, Song, Finin, and Tseng (2009) report half of the tweets that they harvested had some geolocation information, in the form of coordinates or description, while Cheng, Caverlee, and Lee (2010) report that 21% of the tweets in their data corpus had toponym content, while another 5% of them had precise coordinates. While the exact percentage of geolocated tweets may vary depending on the theme and the studied area, we have observed in prior work that on average the percentage of precisely geolocated tweets tends to be much lower, often in the range of 0.5–3.0%.

The ability to convey (and share) thematic as well as location information through social media enables its users to establish *presence* within their domain of interest. While the notion of presence

Please cite this article in press as: Croitoru, A., et al. Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems* (2014), http://dx.doi.org/10.1016/j.compenvurbsys.2014.11.002