A Cyber-archaeology Approach to Social Movement Research: Framework and Case Study

D. Zimbra

Artificial Intelligence Lab, Department of MIS, University of Arizona

A. Abbasi

Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee

H. Chen

Artificial Intelligence Lab, Department of MIS, University of Arizona

This paper presents a cyber-archaeology approach to social movement research. The approach overcomes many of the issues of scale and complexity facing social research in the Internet, enabling broad and longitudinal study of the virtual communities supporting social movements. Cultural cyber-artifacts of significance to the social movement are collected and classified using automated techniques, enabling analysis across multiple related virtual communities. Approaches to the analysis of cyber-artifacts are guided by perspectives of social movement theory. A case study on a broad group of related social movement virtual communities is presented to demonstrate the efficacy of the framework, and provide a detailed instantiation of the proposed approach for evaluation.

Since the 1990s, the Internet has been instrumental in the advancement of many social movements. An early and well-cited example of the effective use of computer-mediated communication (CMC) for activism and protest is the case of the Zapatista movement and the Zapatista National Liberation Army (EZLN) in their rebellion against the ongoing colonial repression of the indigenous people in Mexico (Collier, 1994; Cleaver, 1995). The Zapatista movement leveraged the Internet and CMC "for the dissemination of information, the sharing of experience, and the facilitation of discussion and organizing . . . (and) for the amplification and archiving of the developing history of the struggle" (Cleaver, 1999). Through their use of the Internet and CMC, the Zapatista movement demonstrated "a new capacity for this and other social movements to communicate across borders and to operate at a transnational level" and was considered a "prototype" for other movements (Cleaver, 1999). Related antiglobalization demonstrations inspired by the Zapatistas such as the

J18 and the protests against the World Trade Organization in Seattle during 1999 similarly utilized the Internet and CMC to promote and coordinate their activism, and were also regarded as models of demonstration to be followed in subsequent years (Starr, 2000).

Recognizing the importance of CMC to social movements, researchers have focused studies specifically on the use of the Internet by various movements, from grassroots activism (Tesh, 2002) to transnational mass protest movements (Clark and Themudo, 2006). Social movement activism mediated by the Internet is often referred to cyber-protest, and represents an emerging field of social movement research (Meikle, 2002; Jordan and Taylor, 2004; van de Donk et al, 2004). Cyber-protest is considered an extension of a social movement into a new media space; an articulation of the movement in the Internet (Castells, 1998). Cyber-protest is diverse, and many forms have been identified in the literature, from hactivism—electronic civil disobedience, as in the case of the 1998 blockade of the Pentagon website by the Electronic Disturbance Theater (EDT)—to culture jamming—tactical media using techniques of appropriation, collage, ironic inversion, and juxtaposition, as in the case of the ®™ark websites protesting the Bush presidential campaign of 2000 (Meikle, 2002). Although the Internet has empowered social movements to produce novel expressions of protest, they continue to rely upon the traditional tactics proven effective in the pre-Internet world adapted to a virtual environment (Meikle, 2002). For example, the EDT blockade was regarded by researchers as a virtual sit-in, a traditional and successful method for protest. Following the Zapatista model, the most abundant forms of cyber-protest leverage the Internet as a platform for communication, to disseminate information, share experiences, develop solidarity and identity, and promote, facilitate, and organize activism. The extensive usage of CMC by a social movement creates a widespread cyber-culture in the Internet, establishing many linked virtual communities for members to congregate and socialize (Meikle, 2002), forming rich and complex social systems (Jones, 1994), and representing fertile areas for conducting social research.

While researchers have recognized the importance of the Internet to social movements and the value of studying virtual communities and cyber-protest to further their understanding, they have also identified challenges in conducting social research online. Manifestations of social movements in the Internet and their expressions of cyber-protest are dynamic, appearing and disappearing without warning, and distributed across multiple related websites, forums, or virtual communities, some of which may not be readily apparent to the researcher.

> Although they may include formal organizations as components, on the whole they are not an organization. A social movement typically lacks membership forms, statutes, chairpersons, and the like. It may expand or shrink considerably over periods of time, and exhibit phases of visibility and latency. Also, unlike political parties, social movements may have significant overlaps with other

movements. Moreover, a social movement may quickly change its forms, strategy, tactics, and even some of its goals'' (van de Donk et al., 2004).

The cyber-protest of social movements is ''fuzzy and fluid phenomena often without clear boundaries'', and a ''moving target, difficult to observe'' (van de Donk et al, 2004). The dynamic and open nature of social movements in the Internet, as described by cyber-protest researchers, presents unique challenges to conducting social research online.

Researchers have proposed approaches to social research in the Internet, specifically attempting to address the challenges presented by the scale and complexity of the online environment. Paccagnella suggests that scholars conducting social research on virtual communities ''exploit the possibilities offered by new, powerful, and flexible tools for inexpensively collecting, organizing, and exploring digital data'' (Paccagnella, 1997). Such tools would facilitate the analysis of 'cultural cyber-artifacts,' the member communications and creations representing the archived history of the virtual community. Jones (1997) presents a theoretical outline for conducting social research on virtual communities, adapting the perspectives of archaeology to CMC. Similar to Paccagnella, the approach focuses on the cultural cyber-artifacts of a virtual community, which provide an integrated framework for analysis to understand the life of the community and its members (Jones, 1997). Due to its focus on the study of cultural cyber-artifacts, the approach was called cyber-archaeology. Focusing on the cyber-artifacts of virtual communities and developing sophisticated tools for their collection and analysis, issues of scale and complexity surrounding cyber-protest may be resolved, enabling broad and longitudinal research on social movements. However, few approaches have been proposed in the literature with the technical sophistication and capabilities to assist in social research on modern virtual communities and forms of cyber-protest, and social researchers continue to be slow to react to social movements as they emerge (Edelman, 2001).

In this paper, we present a cyber-archaeology framework for social movement research. We refer to the approach as cyber-archaeology, following Jones (1997), because of our focus on the cultural cyber-artifacts of cyber-protest. Our framework overcomes many of the issues of scale and complexity facing social research in the Internet, and specifically addresses many of the suggestions of Paccagnella (1997) enabling broad and longitudinal social research on virtual communities. The framework is not intended to replace the social researcher, or directly produce theoretical conclusions regarding social movements, but instead we offer a collection of powerful tools and techniques to assist in conducting social movement research in the Internet. The approach is automated and scalable, allowing a researcher to easily collect and analyze social movement cyber-artifacts across multiple linked virtual communities. The techniques for the analysis of cyber-artifacts specifically adopt perspectives of social movement theory. To demonstrate the cyber-archaeology framework for social movement research, and provide a detailed instantiation of the

proposed approach for evaluation, a case study on a broad group of related social movement virtual communities is presented.

Literature Review

Social Movement Theory

Social movements are informal networks formed through the shared beliefs and solidarity of members, which mobilize to support specific positions on social issues through various forms of protest (Della Porta and Diani, 1999). Researchers have studied social movements since the early 20th century, and theories have emerged and developed along two historical lineages: the European and the United States perspectives. The European perspective was founded in Marxism and focused on the fundamental conflict between the proletariat and the bourgeoisie, and the 'historical mission' of the proletariat to overthrow the bourgeoisie through revolution. The early U.S. perspective on social movements is referred to broadly as collective behavior (Park, 1967; Smelser, 1962), has roots in functionalist social psychology (Parsons, 1951), and considers movements as spontaneous and irrational responses to the structural strains and deprivation imposed on members by society.

Following an intense period of activism in the 1960's, both the European and U.S. perspectives went through theoretical transformations during the 1970's. The European perspective on social movement theory produced the new social movement (NSM) approach. The new social movements focused on by NSM researchers included environmentalism, peace, second-wave feminism, and animal rights, among others. NSM theorists diverged from their historical Marxist roots, and adopted "a view that contemporary western societies have outgrown the model of capitalist society suggested by Marx, rejecting the priority he affords to class struggle and to classes as agents of historical change" (Crossley, 2002). They "oppose economic reductionism and class reductionism: the emergence of social movements cannot be explained solely by economic changes and the position of actors in the production process" (Fuchs, 2006a). Instead, NSM researchers stress nonclass issues in their analysis of social movements, such as gender, ethnicity, age, and environment. NSM researchers also focus as much on "the constitutive structure and type of society in which modern movements emerge, the relation of those movements to that society, and their 'historical role' therein" as they do the movement itself, and "attempt to identify the central conflicts and movements (that are) definitive of the new era" (Crossley, 2002). Theorists central to the NSM approach include Habermas (1981), Touraine (1981), and Melucci (1986).

The U.S. perspective also transformed theoretically during the 1970's, and produced the resource mobilization theory (RMT). RMT developed from rational actor theory (RAT) and considered participants in social movements as actors who "rationally calculate gains and losses that stem from certain potential actions and make choices for or against certain actions based on such calculations" (Fuchs, 2006a). The rational actor represented a dramatic departure from collective behavior approaches

that considered social movements as irrational reactions to grievances or deprivation. RM theorists viewed social movements as rational, organized manifestations of collective action and contention, with specific resources mobilized strategically to accomplish their respective objectives (Jenkins, 1983; McCarthy and Zald, 1977). Movement resources included material, human, cognitive, technical, and organizational capital and capabilities, which were utilized to coordinate activism, encourage participation, or gain political power (Klandermans, 1984). RMT also included a community of researchers who adopted a structural 'network' model of social relations and life, which allowed resources to be mobilized throughout the movement (Crossley, 2002; Diani and McAdam, 2003). The RM approach also considers other factors influencing the development of social movements such as framing (Snow and Benford, 1988; Benford and Snow, 2000; McAdam et al., 1996) and political opportunities (McAdam et al., 1996; Tarrow, 1998).

Social theorists have been critical of both the NSM and RMT approaches, scrutinizing their deficiencies. We summarize the general concerns as follows. While the NSM approach focuses on macroaspects of social movements, such as the conditions from which the movement developed and the role of the movement in the history of society, it neglects to address the process by which movements achieve their social and political objectives. On the other hand, the RMT approach focuses on microaspects of social movements, such as the organizational structure and strategies for protest, while neglecting the reasons why social movements form and mobilize. Concisely, NSM explains 'why' while RMT explains 'how.' Crossley (2002) refers to the deficiencies in the NSM and RMT approaches as failures to explain aspects of social movements pertaining to agency, structure, or both. For a detailed critique of the NSM approach, see Hannigan (1985) or Pichardo (1997); for the RMT approach, see Crossley (2002).

Social Movement Theory and the Internet
Since the effective cyber-protest of the Zapatista, social movements and the Internet have developed an intimate relationship (Cleaver, 1995; Clark and Themudo, 2006), even influencing the evolution of each other (Fuchs, 2006b). The inherent structure of the Internet as a network encouraged researchers to reexamine the network model of social movements (Diani and McAdam, 2003). The network model diverged from the more traditional RMT perspective on social movements as organizations with formal, hierarchical, and bureaucratic structures. The new social movements that inspired the network model did not require membership, were decentralized, dynamic, and without formal hierarchy, and depended upon participants identifying with the perspectives and positions of the movement and its objectives (Diani, 2003). For proponents of the model, networks were significant in social movements and fulfilled many required functions. "Networks undoubtedly facilitate mechanisms like the mobilization and allocation of resources across an organizational field, the negotiation of agreed goals, the production and circulation of information, all activities which are also essential to any type of coalition, broadly defined; at the same

time, however, they also may—or may not—facilitate the circulation of meaning and mutual recognition'' (Diani, 2003). The network model of social movements was attractive to cyber-protest and social movement researchers due in part to the structural characteristics of the Internet and CMC. Social movements articulated online are fluid and dynamic networks, maintaining the linked structure of the Internet through their mobilization and expression of cyber-protest. ''In defining social movements, most observers would agree that they need some structure and are to some extent 'organized.' This structure, however, is not an organization but a network, or even a network of networks that both rests on and maintains a sense of collective identity'' (van de Donk et al, 2004).

Diani was one of the early and key contributors to the network model of social movements (1990, 1992), and recognized the potential to extend this model to CMC (Diani, 2000). In (2003), Diani proposes a comprehensive research approach to the study of social movement networks. In introducing the approach, the author reviews the literature and development of the network model of social movements, and describes how taking a network perspective of social movements can specifically improve the theories and research based on RMT, NSM, and political process theories (PPT). The author also provides a detailed description of the proposed approach. The author emphasizes two key characteristics to network structure to guide analysis: centralization and segmentation. Centralization refers to the formality of the network structure, while network segmentation refers to the ability and extent to which actors in the network can communicate without impediment. The author also identifies and describes four types of network structures: clique, policephalous, wheel/star, and segmented and decentralized. The proposed research approach to studying social movement networks consisted of four forms of network analysis: multilevel analysis, link multiplicity, time, and structural homophily. These forms of analysis represent an integrated framework to guide future research on social movement networks.

Fuchs extends the network concept to describe social movements as self-organizing systems (2006a). By adopting the perspective of self-organization and viewing social movements as dynamic and complex systems on both the micro and macrolevels, the author addresses some major theoretical concerns with the RMT and NSM approaches. The author defines self-organizing systems as ''complex networks of entities that synergetically interact and produce novelty'' (Fuchs, 2006a). Through this perspective, the author connects the proposed approach to the literature on the network model of social movements. In the author's view, social movements are ''dynamic communication systems that permanently react to political and societal events with self-organized protest practices and protest communications that result in the emergence and differentiation (production and reproduction) of protest structures (events, oppositional topics, alternative values, regularized patterns of interaction and organization)'' (Fuchs, 2006a). The Internet itself serves as a dynamic communication system, and is self-organized, alluding to the tight integration between the self-organization model of social movements and cyber-protest. This

relationship is described in detail in (Fuchs, 2006b). To the author, "cyber-protest is a global structural coupling and mutual production" of the self-organization of the Internet and the self-organization of social movements (2006b). In the case of the self-organization model, the Internet and cyber-protest have demonstrated strong influence over the development of social movement theory.

Social Research on Virtual Communities

Although we have shown that the Internet has had significant influence on social movements, and has been engaged by researchers and theorists, there are additional characteristics of cyber-protest that make conducting social movement research online appealing. The widespread use of CMC by social movement participants produces rich and complex social systems, such as virtual communities or cyber-societies (Jones, 1994). They are often vibrant with strong membership and robust communication, containing discourse that is highly revealing of the perspectives and ideologies of the social movement. There are both technical and social reasons why the study of virtual communities can benefit social research. Inherent technical features of cyber-protest and virtual communities offer aspects supporting analysis not readily found in natural settings. In a virtual community, all cultural artifacts produced by members are recorded and preserved, and can be recovered by the social scientist if and when desired. Advancement of CMC technology and the emergence of web 2.0 have provided a sufficient level of sophistication for the creation of virtual communities that can approximate real-life social interactions, producing a wide variety of artifacts with cultural and social significance. Traditional studies face the difficult task of determining the time associated with collected artifacts, but in an online environment, time is typically recorded precisely for each cyber-artifact. The accessibility of time information allows for analysis of the exact evolution of interactions and the production of cultural artifacts, which can be highly revealing of social behavior.

The technical features provided by virtual communities are not the only benefits of conducting social research in the Internet. Social research focusing on virtual communities is often highly revealing of social behavior due to the phenomena described by the social identity model for deindividuation (SIDE) (Reicher et al, 1995). SIDE evolved in parallel with social movement theory, but differed in the perspectives of individual behavior within groups. Traditional models of deindividuation based on collective behavior describe factors such as the immersion in the group and lack of anonymity that lead to a loss of selfhood and control over individual behavior. However, SIDE theory shows that as deindividuation occurs in communities, individuals conform to categorical norms and organize to maximize the power of their collective identity in social movements (Reicher et al, 1995). Researchers have also studied CMC as a potential channel for social resistance, showing that due to the behavior of individuals described by SIDE, CMC was highly effective for communicating the social support necessary to resist powerful outgroups (Spears et al, 2002). These findings suggest that conducting social movement research focused on cyber-protest

and virtual communities may be even more revealing of the true social behavior of participants than observation in a more natural setting.
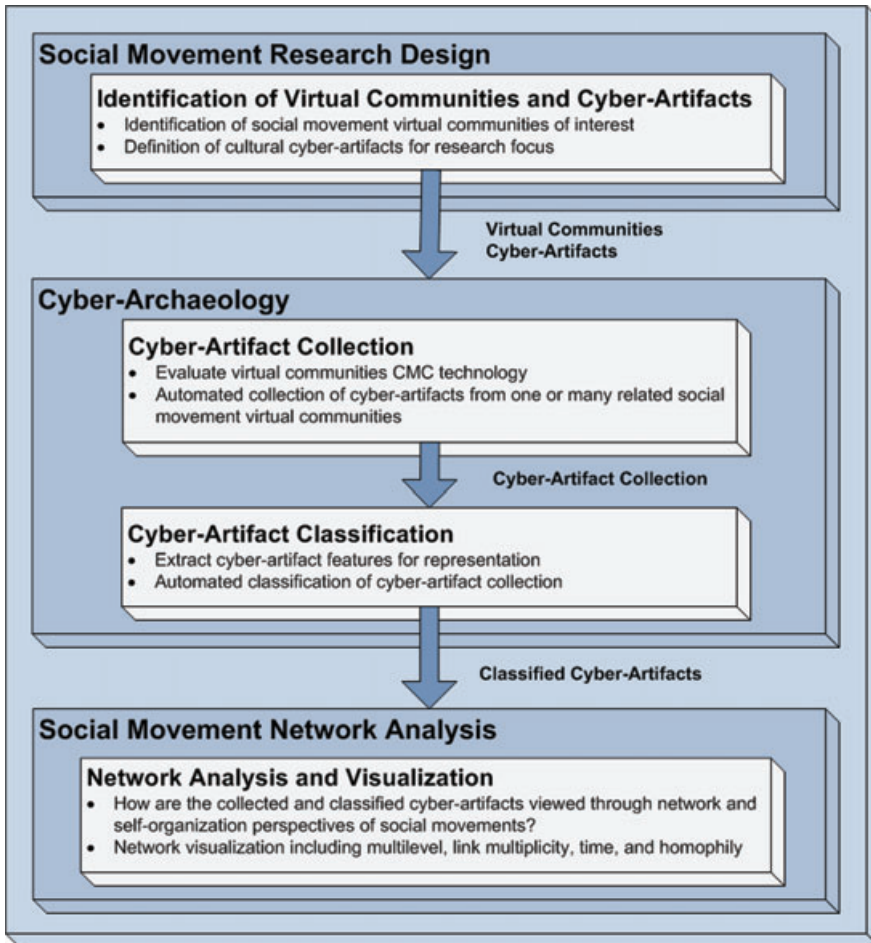
Cyber-archaeology

Many social researchers have recognized the scale and complexity of cyber-protest, and the difficulty of conducting social research in the Internet. Paccagnella suggests that "deep, interpretive research on virtual communities" could be greatly assisted by "new analytic, powerful yet flexible tools" for "collecting, organizing and exploring digital data" (Paccagnella, 1997). Such tools would facilitate the analysis of 'cultural cyber-artifacts,' the member communications and creations representing the archived history of the virtual community (Paccagnella, 1997; Jones, 1997). Cultural cyber-artifacts are diverse and could include websites, forum communications, e-mail, video or audio clips, documents, chat, bulletin board messages, and usage logs. The tasks performed by these tools to assist the social researcher in analyzing the cyber-artifacts of a virtual community would include data collection, preparation, and coding, storage and search, and data linking and display (Paccagnella, 1997). While the author never suggests that these tools could replace the social researcher in the analysis, since "computers don't analyze data, people do," such techniques could greatly enhance the virtual community research and alleviate some of the technical challenges presented by CMC and the Internet environment.

Cyber-Archaeology Framework for Social Movement Research

We present a cyber-archaeology framework for social movement research. Our framework overcomes many of the issues of scale and complexity facing social research in the Internet, and specifically addresses many of the suggestions of Paccagnella (1997) enabling broad and longitudinal social research on virtual communities. The framework is not intended to replace the social researcher, or directly produce theoretical conclusions regarding social movements, but instead we offer a collection of powerful tools and techniques to assist in conducting social movement research in the Internet. The approach is automated and scalable, allowing a researcher to easily collect and analyze social movement cyber-artifacts across multiple linked virtual communities. The techniques for cyber-artifact analysis adopt perspectives of social movement theory, specifically the network (Diani and McAdam, 2003) and self-organization (Fuchs, 2006a) models of social movements.

The proposed framework for cyber-archaeology has three phases, as presented in Figure 1. In phase one, social movement research design, social researchers identify the social movement of interest, associated virtual communities, and target cyber-artifacts for collection and analysis. Phase two consists of the automated collection and classification of cyber-artifacts across one or many identified virtual communities. The third phase supports the analysis of cyber-artifacts by social researchers from the perspectives of social movement theory, including various network analysis and visualization techniques to perform multilevel, link, time, and homophily analysis, as suggested by Diani (2003). Each of these phases is discussed in detail below.

Figure 1 Cyber-Archaeology Framework

The first phase of the cyber-archaeology framework consists of the social research design on the movements of interest. Researchers identify a social movement of interest and their related virtual communities, in consultation with domain experts when necessary, to develop a seed set of websites for analysis. The list of targeted virtual communities can then be expanded through link analysis of the seed set of sites, to include websites linked to and from the virtual communities already identified. Once a satisfactory list of virtual communities has been defined, the targeted cultural cyber-artifacts for research focus must be determined. A lexicon of terms pertaining to the cyber-artifacts of interest must also be developed to assist in the collection process by enabling the identification of relevant content. By limiting the search to specific cyber-artifacts of interest, the social researcher can focus their investigation, thereby alleviating issues of scale impacting social research in the Internet. The cyber-artifacts selected could be based upon, for example, influential

figures in the movement, meaningful events, or core ideologies and symbolism. Focusing on specific cyber-artifacts that develop the identity of the movement and create solidarity among members deepens the social research. However, a researcher need not restrict their focus to specific cyber-artifacts; the framework supports a more comprehensive approach. Without restriction, all cyber-artifacts of the virtual communities may be collected for analysis, potentially overwhelming the researcher, but providing completeness in the research.

The second phase consists of the automated collection and categorization of the targeted cyber-artifacts of interest from the virtual communities identified in the first phase. We consider the collection and classification tasks to be core to any study on social movements in the Internet. Researchers have highlighted the promise for enriching the virtual community research field by developing automated techniques to assist with these tasks (Paccagnella, 1997). Researchers have also identified opportunities to improve the understanding of social movements by broadening the research in scope, beyond a single-organization or single-issue focus (Edelman, 2001). To aid in the analysis of the collected cyber-artifacts, automated models for classification can be developed to categorize cyber-artifacts across numerous dimensions depending upon their nature and the goal of the research. Developing automated techniques for cyber-artifact collection and categorization allows social researchers to ask questions that would be infeasible to answer by traditional, manual methods, expanding the scope of research and the understanding of social movements. The outputs of the second phase, the collected and classified cyber-artifacts of the virtual communities, are then analyzed manually by the social researcher to learn about the movement. The third phase of the framework is considered optional, although these analyses are guided by social movement theory. Therefore, phases one and two of the framework could also support other forms of analysis, if desired by the social researcher.

The third phase of the framework consists of the analysis of the collected and classified cyber-artifacts by social researchers from the perspectives of social movement theory. In particular, we adopt the network and self-organization models of social movements. Three forms of network analysis and visualization techniques are presented, providing insight into the patterns of self-organization exhibited by the virtual community with respect to the collected and classified cyber-artifacts. These network analyses and visualizations follow the research program proposed by Diani (2003), and cover multilevel analysis, link multiplicity, the evolution of networks over time, and structural homophily. The methods developed to support the third phase of the framework also utilize structured approaches and automation, similar to the techniques developed in the second phase, to support scalability in the social research.

Case Study

To demonstrate the efficacy of the cyber-archaeology framework, and provide a detailed instantiation of the proposed approach for evaluation, we present a case study on a broad group of related social movement virtual communities.

Introduction

Some of the most sophisticated virtual communities with strongest membership and social ties between community members belong to socially extreme movements. There are many social movements throughout the world that could be considered socially extreme or extremist, struggling for various causes based upon race, gender, sexuality, and religion. Extremist movements actively use the Internet as a quick, inexpensive, and anonymous means of communication, overcoming the limitations of traditional media, and allowing their ideological information to reach new audiences around the world. Studies have shown that these groups promote their positions through their Internet communications and publications (Glaser et al, 2002), and found evidence of resource sharing, fund raising, propaganda, recruitment, and training materials (Burris et al, 2000; Shafer, 2002; Zhou et al, 2005). Social scientists have long been interested in Middle Eastern movements supporting anti-Western positions, some of which contain participants or subgroups with socially extreme perspectives whose activism includes high-risk behavior, applying social movement theory to understand these societies and individuals (Snow and Marshall, 1984; Wiktorowicz, 2002). To be clear, we do not consider all anti-Western movements of the Middle East to be extreme, nor do we believe that all members of such movements engage in extremist behavior in their activism. However, some of the participants in these social movements hold extreme positions and engage in high-risk behavior in their activism, with grave cost. Social researchers have specifically studied the motivations and dynamics of high-risk behavior in social movements (McAdam, 1986). Communities supporting high-risk activism stress the importance of social networks, as strong social ties are required to provide the trust and solidarity among activists to encourage the high-risk behavior (Wiktorowicz, 2002). In studying these social movements with some participants whose activism includes high-risk behavior, we perform a cyber-archaeological analysis on a specific theme of cultural cyber-artifacts significant to some anti-Western Middle Eastern movements: content related to improvised explosive devices (IEDs). IEDs have social significance among some movement participants as an enabler of violent uprising and resistance, equalizer of power, and as the focus of high-risk activism and related symbolism. Examining IED-related cyber-artifacts in a global social context can provide a deeper understanding of these social movements, and the high-risk behavior of some of the participants.

Cyber-Archaeology Framework Phase 1

In the first phase of the cyber-archaeology framework, a list of social movements of interest and related virtual communities is defined. For the case, the social movements and virtual communities were identified using the expertise of a number of government and research organizations. The seed set of social movement virtual communities was expanded through link analysis. In-links to and out-links from the virtual communities were collected and examined manually to identify additional related virtual communities to be included in the social movement research. Utilizing

an automated approach to cyber-artifact collection and analysis, social movement research can extend across multiple related virtual communities. Since the social movements identified for this research communicate primarily in Arabic, a lexicon of terms relevant to IEDs was developed with assistance from domain and language experts to direct the cyber-artifact collection effort. The lexicon for IED-related cyber-artifacts consisted of more than 100 terms and their linguistic variations; a subset of the terms is shown in Table 1 (WWW resource, 2010).

Cyber-Archaeology Framework Phase 2
The second phase of the cyber-archaeology framework consists of the automated collection and categorization of IED-related cyber-artifacts from the identified social movement virtual communities. The automated approach provides the ability to expand the research to multiple virtual communities and acquire a more complete collection of cyber-artifacts of interest.

Cyber-Artifact Collection
The collection of IED-related cyber-artifacts from social movement virtual communities entails the use of focused crawling techniques. Focused crawlers are intelligent web spiders designed to retrieve content relating to a specific subject. They ''seek, acquire, index, and maintain pages on a specific set of topics that represent a narrow segment of the web'' (Chakrabarti et al, 1999). Once the focused crawler has been properly configured, cyber-artifacts of interest can be collected from multiple virtual communities in an automated fashion with little supervision. For the purposes of the case study, these focused crawling techniques were applied in the collection of IED-related cyber-artifacts from 30 virtual communities identified in phase 1. In total, 2,541 IED-related web page cyber-artifacts were collected and qualified through manual inspection by a domain expert. Over 90% of the cyber-artifacts were gathered from a core set of 7 websites, as shown by the frequency distribution in Figure 2 (WWW resource, 2010) and collection summary in Table 2 (WWW resource, 2010).

Cyber-Artifact Classification
To assist in the analysis of the collected cyber-artifacts, automated models for classification can be developed to categorize cyber-artifacts across numerous dimensions depending upon their nature. For CMC communications, a useful categorization for analysis of cyber-artifacts is by genre of communication. Genre is a literary and rhetorical concept that describes widely recognized types of discourse (Yates and Orlikowski, 1992). Genres of communication are traditionally defined in terms of shared purpose and common conventions of substance and form (Herring and Paolillo, 2006). Shared purpose refers to the social motives and intentions behind the communication. Substance refers to the themes and topics expressed in the communication. Form refers to observable physical and linguistic features of the communication, including structural attributes, communication medium, and language such as degree of formality and use of a specialized vocabulary. Genres of

Table 3 Description of the Genres of IED Web Pages

Genre	Type	Description
Material Pages	Weapons	Detailed descriptions of various *weapons and IEDs*. Includes training materials on how to build, use, and maintain devices.
	Attacks Reports	Reports on the *occurrence of an attack* with some elaboration on the specific details of the operation.
	Military and Recon Reports	Reports on the latest *military intelligence* (provides statistics of operations, troop and vehicle movements, etc.).
	Tactics	Discusses the pros and cons of various *military tactics* and provides examples from previous operations.
Discussion Pages	General Discussion	General discussion of *IEDs* and *IED related events*.

communication are used within a community to establish expectations about the purpose, content, participants, form, time, and place of the interaction, describing the who, what, where, when, why, and how of the communication (Yates and Orlikowski, 2002).

Traditionally, researchers have analyzed genres to understand the communication dynamics of a community or organization. Yates and Orlikowski (1992) observed that the emergence of genres within an organization is due to particular sociohistorical contexts, which are reinforced over time as the situation recurs. Researchers have also developed techniques for the automated classification of documents by genre. While genre classification is similar in nature to topic or author classification, the representative features used to characterize the document differs according to the type of classification being performed. In topical classification, features are dependent upon the topic, whereas in author and genre classification features are often considered orthogonal to topic, and greater emphasis is placed on the stylistic and structural elements of the document.

Classification of collected CMC cyber-artifacts by genre of communication assists in the analysis of the social movement virtual community. Through manual analysis of the collected IED-related CMC cyber-artifacts by a domain expert, two distinctive genres of communication were identified: web pages containing descriptions on IED-related weapons, attacks, and military tactics (materials pages), and web pages with general discussion of IEDs or IED related events (discussion pages). The two genres of web pages identified are described briefly in Table 3. Sample web pages from each type are presented in Figures 3–7 (WWW resource, 2010).

Developing automated methods for the classification of these cyber-artifacts provides the ability to enhance and expand the scope of the social research on

movement virtual communities. Our genre classification approach for IED-related web pages is comprised of an extended feature representation coupled with machine learning algorithms. The extended feature representation incorporates rich structural information with stylistic and topical linguistic features, intended specifically to represent the purpose, substance, and form of the web page, indicative of genres of communication (Yates and Orlikowski, 2002). Structural features include HTML tags and information about the technical structure of the web page, which can specifically express genre information to differentiate materials and discussion pages. Stylistic features include word- and character-level lexical measures, word length distributions, vocabulary richness, function words and punctuation usage, and effectively capture the stylistic tendencies of the web page (Abbasi and Chen, 2005; Zheng et al., 2006). Topical features such as word n-grams are used to identify themes and topics in web pages. Only features exceeding a threshold number of occurrences (3) were included in the representation. The extended feature set is summarized in Table 4.

The support vector machine (SVM) learning algorithm was used with the extended feature representation to automatically classify the collected IED-related web pages. The SVM is a statistical based algorithm that has been used successfully in text categorization (Joachims, 1998), and is particularly suited to learning with extended feature representations. Classification models that utilize extended feature representations often leverage methods for feature selection to identify the most relevant features for inclusion to the model, thereby eliminating unnecessary information and reducing model complexity. For the classification of the IED-related web page cyber-artifacts, we evaluate two SVM models: The first included the entire extended feature set (labeled SVM), while the second used feature selection on the extended feature set based upon the information gain heuristic, a measure of entropy (labeled SVM-IG). Both SVM classification models utilized linear kernel functions.

The test bed of 2,541 collected IED-related cyber-artifact web pages was used for the development and evaluation of the automated genre classifier. As previously stated, 2,501 discussion and 40 materials web pages were identified through manual categorization by a domain expert. Due to the large number of discussion pages relative to materials pages, adjustments were made to provide equal representation of materials and discussion pages in the classification experiments. 100 iterations of the experiment were performed, and in each iteration 40 discussion pages were randomly selected from the 2,501 to complement the 40 materials web pages; a different set of 40 discussion pages were used in each iteration of the experiment. 10 fold cross validation was performed to test classification and generate performance results, measured by computing the percentage of correctly classified web pages. Overall performance values were calculated as simple averages of the 100 iterations.

Overall, both the SVM and SVM-IG models performed well in genre classification of IED related materials and discussion web pages, with better than 81% accuracy. The SVM-IG model using the extended feature set and the information gain heuristic for feature selection performs particularly well with over 88% accuracy, and outperforms

Table 4 Extended Feature Set for Genre Classification

Group	Category	Quantity	Description/Examples
Stylistic	Word-Level Lexical	5	total words, % char. per word
	Character-Level Lexical	5	total char., % char. per message
	Character N-Grams	<18,278	count of letters, char. bigrams, trigrams (e.g., كك,اب)
	Digit N-Grams	<1,110	count of digits, digit bigrams, digit trigrams (e.g., 1, 12, 123)
	Word Length Dist.	20	frequency distribution of 1−20 letter words
	Vocabulary Richness	8	richness (e.g., hapax legomena, Yule's K, Honore's H)
	Special Characters	21	occurrences of special characters (e.g., @# $%^& ∗ + =)
	Function Words	300	frequency distribution of function words (e.g., of, for, to)
	Punctuation	12	occurrence of punctuation marks (e.g., !; :, .?)
	Word Root N-Grams	varies	roots, bigrams, trigrams (e.g., كسب, كتب)
Structural	Message-Level	6	e.g., has greeting, has url, requoted content
	Paragraph-Level	8	e.g., number of paragraphs, sentences per paragraph
	Technical Structure	50	e.g., file extensions, fonts, use of images, HTML tags
	HTML Tag N-Grams	<46,656	e.g., <head>, , <td>, <message>
Topical	Word N-Grams	varies	bag-of-words n-grams (e.g., "explosive", "explosive device")

the standard SVM model by 7%. A t-test was performed to determine whether the performance increase gained by utilizing feature selection was significant. The performance of the SVM-IG model were a statistically significant improvement ($p < 0.001$) over the model that used the entire feature representation without feature selection. Furthermore, the SVM-IG model was more reliable, performing with better than 80% classification accuracy in every iteration of the experiment. The information gain heuristic identified 9,268 of the 21,333 features as key and relevant to the genre classification. The results of the experiments are presented in Table 5 and Figure 8 (WWW resource, 2010).

Cyber-Archaeology Framework Phase 3
The third phase of the cyber-archaeology framework consists of the analysis of collected IED-related cyber-artifacts through the perspectives of social movement

**Table 5** Results of Genre Classification of IED Web Pages

Model	Features	Mean Accuracy	Standard Deviation	Range
SVM	21,333	81.938%	5.313%	65.0%−92.5%
SVM-IG	9,268	88.838%	3.238%	80.0%−96.3%

theory. In particular, we adopt the network and self-organization models of social movements. Three forms of network analysis and visualization techniques are presented, providing insight into the patterns of self-organization exhibited by the virtual community with respect to the collected and classified cyber-artifacts. These network analyses and visualizations follow the research program proposed by Diani (2003), and cover multilevel analysis, link multiplicity, the evolution of networks over time, and structural homophily. The methods developed to support the third phase of the framework also utilize structured approaches and automation, similar to the techniques developed in the second phase, to support scalability in the social research.

In the case of IED-related cyber-artifacts, learning how the highly specialized knowledge pertaining to these technologies is organized, distributed, and leveraged throughout the related social movement virtual communities is of particular research interest. Furthermore, an understanding of the communication patterns related to these cultural artifacts is significant to the understanding of the social movements, and the relations the communities have with IEDs and related high-risk activism.

The third phase of the framework consists of three forms of network analysis and visualization, addressing multilevel analysis, link multiplicity, time, and homophily (Diani, 2003). The first type of network analysis, site map analysis, visualizes the self-organization of discussions and resources pertaining to the cyber-artifact of interest within virtual communities. Site map analysis addresses the multilevel analysis suggested by Diani (2003). The second type of network analysis, link analysis, visualizes the ways virtual communities interact, self-organize, and share resources. This form of network visualization addresses the link multiplicity analysis proposed by Diani (2003). The third type of network analysis visualizes the growth of virtual communities over time, addressing the temporal analysis suggested by Diani (2003). By leveraging the classification of the cyber-artifacts performed in the second phase, the three forms of network analysis and visualization can also provide insight into the structural homophily of the various classes of cyber-artifacts in the virtual communities. Comparing the classes of collected cyber-artifacts, patterns and differences in structural homophily may be revealed in the three forms of network analysis.

The first type of network analysis, site map analysis, determines how IED-related cyber-artifacts are organized within a single virtual community domain, providing a form of multilevel analysis (Diani, 2003). The average site maps for websites containing IED-related cyber-artifacts were constructed to analyze the intensities of discussion and materials web pages at various levels within the domain, shown in Figures 9 and 10, respectively. The site map was generated by counting the frequency
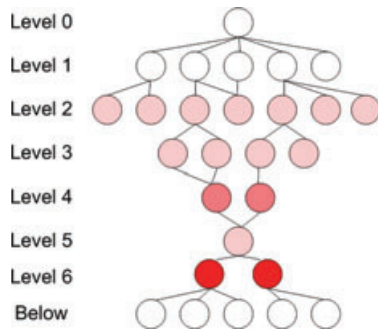
Figure 9 Site Map Featuring Intensity of Materials Pages in Page Levels

of pages for each level and dividing by the total number of web sites. These site maps are indicative of the communication patterns of the participants, showing where within their virtual community these issues are being discussed and specific IED-related resources mobilized. The site maps were drawn to a 20:1 scale; thus, 5 level 1 pages on the map represents 100 actual pages. The intensities of discussion and materials web pages at various levels of the site map are represented using different color shades, with dark colors being more intense concentrations. The distribution of discussion and materials pages among the page levels in the average site maps are also presented in Table 6 (WWW resource, 2010).

The majority of materials web pages occur in domain levels 2–4, while discussion content is often located in levels 2–6. The occurrence of IED-related cyber-artifacts at the middle levels of the domain may indicate that authors do not want their content to be too readily visible, but also do not want to bury their content too deep within the domain as to not be seen by community members. The concentration of discussion pages at lower levels in the domain than materials pages may be attributed to the strong identity traces in discussion content left by the message author, which the author wishes to hide. Materials pages are usually formally written manuals, often in document format, while discussion pages are casual free text with the author's stylistic tendencies evident. These observed differences between the organization of
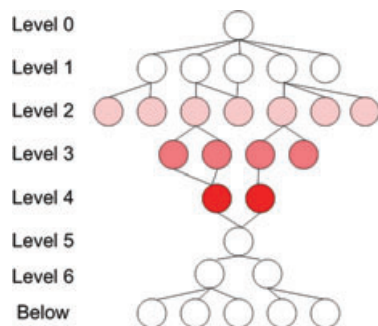

Figure 10 Site Map Featuring Intensity of Discussion Pages in Page Levels
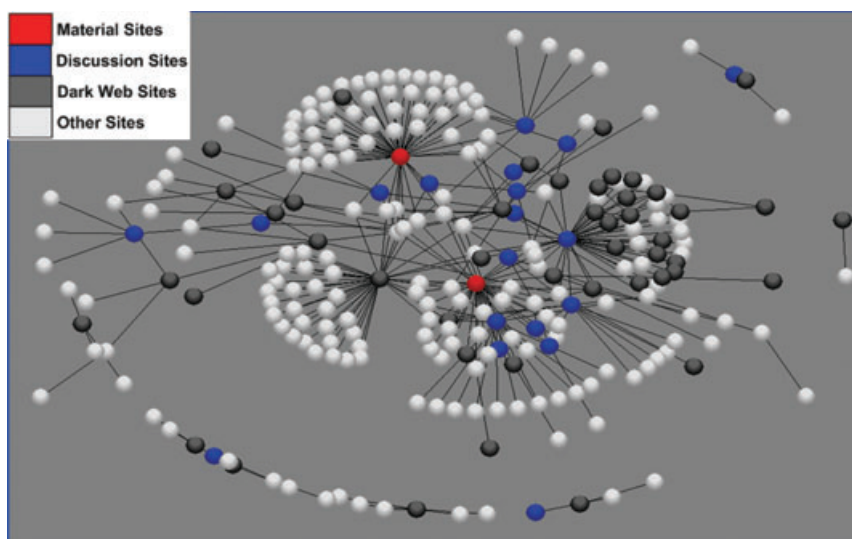
Figure 11 Link Network Structure of Social Movement Virtual communities

materials and discussion pages may be indicative of aspects of structural homophily (Diani, 2003). The site map analysis also identified key terms in the IED lexicon, and their occurrence distribution across site levels. Results revealed that certain terms occur more frequently at specific site levels. A selection of terms is presented in Table 7 (WWW resource, 2010). Also shown are the common levels where the terms are likely to occur, with their average occurrence per page.

The second type of network analysis, link analysis, determines how multiple related virtual communities interact, share cyber-artifacts, and develop their collective identity. This visualization addresses the suggestion of Diani (2003) for analysis of link multiplicity. The link network structure, as displayed in Figure 11, was constructed to show how virtual communities containing IED-related cyber-artifacts were connected, organizing their resources across multiple related sites of the social movement. Each node of the link network represents a website, and is color coded depending upon the IED-related cyber-artifacts collected from the virtual community. The link network also includes websites not previously analyzed in the social movement study but with linkages to the investigated virtual communities. These websites, particularly those with extensive linkages to multiple virtual communities from which IED-related cyber-artifacts were already collected, are strong candidates for future related studies as their linkages may represent the presence of IED-related cyber-artifacts of interest. Three major hubs for materials related IED cyber-artifacts were identified in the network map. As shown in the link network, these websites act as major hubs for the mobilization of IED-related resources. The link network map provides insight into the relationship between virtual communities, and how the production and reproduction of cyber-artifacts occurs between related constituents (Fuchs, 2006b).
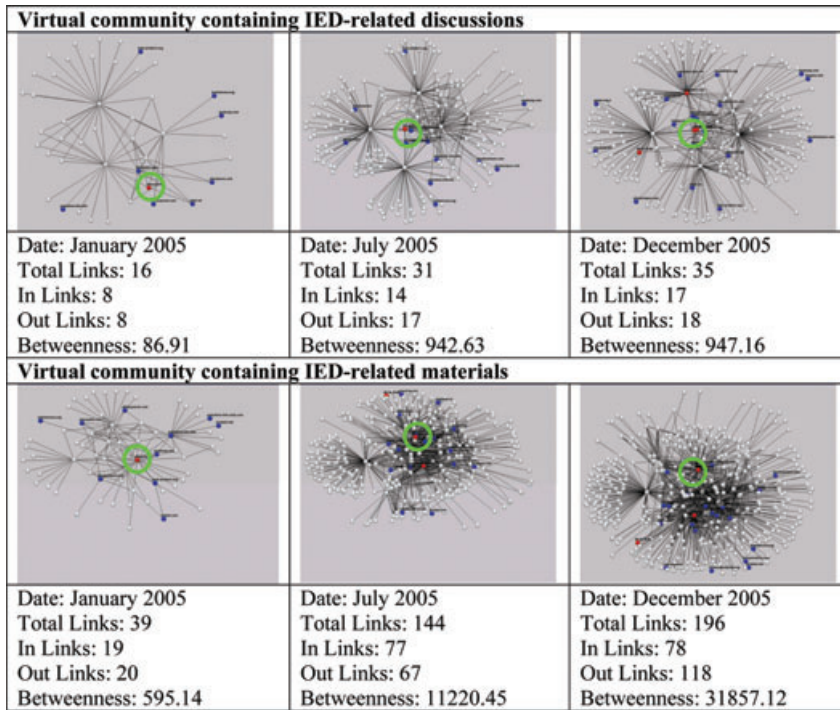
Figure 12 Temporal Views of Material and Discussion Website Link Networks

To understand the evolution of the virtual communities and their self-organization over time (Fuchs, 2006b), and address the temporal dimension of network analysis proposed by Diani (2003), link network snapshots were taken over a 1-year period for two major virtual communities, presented in Figure 12. In the link network diagrams, the virtual community websites containing IED-related discussions and materials, respectively, are circled in green. Three snapshots were taken during the one year period, in January, July, and December of 2005. Both virtual communities experienced significant growth in their link networks over that time. In the case of the discussion website, total links to and from the website increased from 16 to 35. Diani suggested one of the significant aspects of social movement networks was centralization (2003). Betweenness is a measure of centrality in a link network, representing the degree to which a node provides a path to information resources to the rest of the network. The betweenness of the discussion website increased from 86.91 to 947.16, signifying an expanded role as information broker to the network. The link network of the virtual community website containing IED-related materials, expanded dramatically in comparison. Total links to and from the materials website increased from 39 to 196, and the website's betweenness measure increased from 595.14 to 31,857.12. The explosion in the materials website link network may be due to the specialized knowledge contained in IED-related materials cyber-artifacts. These resources are considered particularly valuable to the virtual communities, and are

mobilized quickly and diffused throughout the movement. The differences observed in the growth of the link network supporting the virtual communities containing discussion and materials pages is revealing of structural homophily (Diani, 2003). Temporal analysis of the virtual community link networks provides specific insight into the growth and integration of the community in the larger movement, the production and reproduction of movement participation, and the emergence and distribution of its specialized resources (Fuchs, 2006b).

Conclusion

The Internet, virtual communities, and expressions of cyber-protest are significant to social movements, and offer ideal environments for conducting social research. In this paper, we present a cyber-archaeology framework for social movement research. Our framework overcomes many of the issues of scale and complexity facing social research in the Internet, and specifically addresses many of the suggestions of Paccagnella (1997) enabling broad and longitudinal social research on virtual communities. We offer a collection of powerful tools and techniques to assist in conducting social movement research in the Internet. The approach is automated and scalable, allowing a researcher to easily collect and analyze social movement cyber-artifacts across multiple linked virtual communities. The techniques for cyber-artifact analysis adopt perspectives of social movement theory, specifically the network (Diani and McAdam, 2003) and self-organization (Fuchs, 2006a) models of social movements.

To demonstrate the cyber-archaeology framework for social movement research, and provide a detailed instantiation of the proposed approach for evaluation, a case study on a broad group of related social movement virtual communities is presented. Utilizing focused crawling, cyber-artifacts of interest were collected from multiple linked virtual communities. To categorize the collected cyber-artifacts by genre of communication, an extended feature representation and machine learning models were developed and tested, with nearly 90% accuracy in genre classification. Finally, the cyber-artifacts were analyzed through perspectives of self-organization and the network model of social movements, using network analysis and visualization techniques.

We hope the case study has demonstrated the efficacy of the proposed framework for social movement research, and promotes its future use by researchers. Utilizing the proposed automated approaches to cyber-artifact collection and classification allows for large-scale social research across a greater number of virtual communities compared to traditional methods. A cyber-archaeology approach guided by social movement theory can produce high-quality social research and a deeper understanding of widespread movements.

Acknowledgements

Collection and Infrastructure for Computational and Social Sciences," 2007–2010; and (2) NSF, "EXP-LA: Explosives and IEDs in the Dark Web: Discovery, Categorization, and Analysis," 2007–2010.

References

Abbasi, A. & Chen, H. (2005). Identification and comparison of extremist-group web forum messages using authorship analysis. *IEEE Intelligent Systems*, 20(5), 67–75.

Benford, R. & Snow, D. (2000). Framing processes and social movements: an overview and assessment. *Annual Review of Sociology*, 26, 611–639.

Burris, V., Smith, E., & Strahm, A. (2000). White supremacist networks on the Internet. *Sociological Focus*, 33(2), 215–235.

Chakrabarti, S., Van Den Berg, M., & Dom, B. (1999). Focused crawling: A new approach to topic-specific resource discovery. In *Proceedings of the Eighth WWW Conference.* Toronto, Canada.

Castells, M. (1998). *End of millennium — The information age: Economy, society, and culture.* Blackwell Publishers.

Collier, G. (1994). *Basta! Land and the Zapatista rebellion in Chaipas.* Food First Books.

Clark, D. & Themudo, N. (2006). Linking the web and the street: Internet-based 'dotcauses' and the 'anti-globalization' movement. *World Development*, 34(1), 50–74.

Cleaver, H. (1995). The Zapatistas and the electronic fabric of struggle. Online, http://www.eco.utexas.edu/faculty/Cleaver/zaps.html.

Cleaver, H. (1999). Computer-linked social movements and the global threat to capitalism. Online, http://www.eco.utexas.edu/faculty/Cleaver/polnet.html.

Crossley, N. (2002). *Making sense of social movements.* Open University Press.

Della Porta, D. & Diani, M. (1999). *Social movements: An introduction.* Blackwell.

Diani, M. (1990). The network structure of the Italian ecology movement. *Social Science Information*, 29(1), 5–31.

Diani, M. (1992). Analyzing movement networks. In Diani, M. & Eyerman, R. (eds.), *Studying Collective Action*, Sage.

Diani, M. (2000). Social movement networks virtual and real. *Information, Communication, & Society*, 3, 386–401.

Diani, M. (2003). Networks and social movements: A research programme. In Diani, M. & McAdam D. (eds.), *Social Movements and Networks. Relational approaches to collective action.* Oxford University Press.

Diani, M. & McAdam D. (2003). *Social movements and networks: Relational approaches to collective action.* Oxford University Press.

Edelman, M. (2001). Social movements: Changing paradigms and forms of politics. *Annual Review of Anthropology*, 30, 285–317.

Fuchs, C. (2006a). The self-organization of social movements. *Systemic Practice and Action Research*, 19(1), 101–137.

Fuchs, C. (2006b). The self-organization of cyberprotest. In Morgan, K., Brebbia, C., & Spector, J. (eds.), *The Internet Society II.*

Glaser, J., Dixit, J., & Green, D. (2002). Studying hate crime with the Internet: what makes racists advocate racial violence? *Journal of Social Issues*, 58(1), 177–193.

Habermas, J. (1981). New social movements. *Telos*, 49, 33–37.

Hannigan, J. (1985). Alain Touraine, Manuel Castells and social movement theory a critical appraisal. *The Sociological Quarterly*, 26(4), 435−454.

Herring, S. & Paolillo, J. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4), 439−459.

Jenkins, J. (1983). Resource mobilization theory and the study of social movements. *Annual Review of Sociology*, 9, 527−553.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the ECML*, (pp. 137−142).

Jones, Q. (1997). Virtual-communities, virtual settlements, and cyber-archaeology: a theoretical outline. *Journal of Computer-Mediated Communication*, 3(3).

Jones, S. (1994). *Cyber-society: Computer-mediated communication and community.* Sage Publications.

Jordan, T. & Taylor, P. (2004). *Hacktivism and cyberwars: Rebels with a cause?* Routledge.

Klandermans, B. (1984). Mobilization and participation: social-psychological expansions of resource mobilization theory. *American Sociological Review*, 49(5), 583−600.

McAdam, D. (1986). Recruitment to high-risk activism: The case of freedom summer. *The American Journal of Sociology*, 92(1), 64−90.

McAdam, D., McCarthy, J., & Zald, M. (1996). *Comparative Perspectives on Social Movements.* Cambridge University.

McCarthy, J. & Zald, M. (1977). Resource mobilization and social movements: A partial theory. *The American Journal of Sociology*, 82(6), 1212−1241.

Meikle, G. (2002). *Future active: Media activism and the Internet.* Routledge.

Melucci, A. (1986). *Nomads of the present.* Radius.

Paccagnella, L. (1997). Getting the seats of your pants dirty: Strategies for ethnographic research on virtual communities. *Journal of Computer-Mediated Communication*, 3(1).

Park, R. (1967). On social control and collective behavior: Selected papers. In R. Turner, (ed.), University of Chicago Press.

Parsons, T. (1951). *The Social System.* Routledge.

Pichardo, N. (1997). New social movements: A critical review. *Annual Review of Sociology*, 23, 411−430.

Reicher, S., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology*, 6(1), 161−198.

Schafer, J. (2002). Spinning the web of hate: Web-based hate propagation by extremist organizations. *Journal of Criminal Justice and Popular Culture*, 9(2), 69−88.

Smelser, N. (1962). *Theory of Collective Behavior.* Free Press.

Snow, D. & Benford, R. (1988). Ideology, frame resonance, and participant mobilization. *International Social Movement Research*, 1, 197−218.

Snow, D. & Marshall, S. (1984). Cultural imperialism, social movement, and the Islamic revival. In L. Kriesberg (ed.), *Research in Social Movements, Conflicts, and Change, Vol. 7.* London: Jai Press.

Spears, R., Lea, M., Corneliussen, R., Postmes, T., & Haar, W. (2002). Computer-mediated communication as a channel for social resistance: The strategic side of SIDE. *Small Group Research*, 33(5), 555−574.

Starr, A. (2000). *Naming the enemy: Anti-corporate movements confront globalization.* Zed Books.

Tarrow, S. (1998). *Power in movement.* Cambridge University Press.

Tesh, S. (2002). The Internet and the grass roots. *Organization and Environment*, 15(3), 336–339.

Touraine, A. (1981). *The voice and the eye.* Cambridge University Press.

Van de Donk, W., Loader, D., Nixon, P., & Rucht, D. (2004). *Cyberprotest: New media, citizens and social movements.* Routledge.

Wiktorowicz, Q. (2002). Islamic activism and social movement theory: A new direction for research. *Mediterranean Politics*, 7(3), 187–211.

WWW resource (2010). ai.arizona.edu/papers/zimbra/cyberarchaeology/figures/index.asp

Yates, J. & Orlikowski, W. (1992). Genres of organizational communication: A structurational approach to studying communication and media. *Academy of Management Review*, 17(2), 299–326.

Yates, J. & Orlikowski, W. (2002). Genre systems: Structuring interaction through communicative norms. *Journal of Business Communication*, 39(1), 13–35.

Zheng, R., Qin, Y., Huang, Z., & Chen, H. (2006). A framework for authorship analysis of online messages: Writing-style features and techniques. *Journal of the American Society for Information Science and Technology*, 57(3), 378–393.

Zhou, Y., Reid, E., Qin, J., Chen, H., & Lai, G. (2005). U.S. extremist groups on the web: Link and content analysis. *IEEE Intelligent Systems*, 20(5), 44–51.

About the Authors

David Zimbra is a doctoral student in the Department of MIS in the Eller College of Management at the University of Arizona. He received an M.S. and B.S. in Information Systems from Santa Clara University. His research interests include web mining, social media analytics, business intelligence, and time series forecasting.

Address: Department of MIS, Eller College of Management, University of Arizona, Tucson AZ 85721

Ahmed Abbasi is an assistant professor in the Sheldon B. Lubar School of Business at the University of Wisconsin-Milwaukee. He received his Ph.D. in MIS from the University of Arizona and an M.B.A. and B.S. in Information Technology from Virginia Tech. His research interests include development and evaluation of technologies for enhanced analysis of computer-mediated communication.

Address: Department of MIS, Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, Milwaukee WI 53201–0742

Hsinchun Chen is McClelland Professor of MIS at the University of Arizona. He received a B.S. from the National Chiao-Tung University in Taiwan, an MBA from SUNY Buffalo, and a Ph.D. in Information Systems from New York University. His research interests include digital libraries, data/text/web mining, intelligence analysis, and security informatics.

Address: Department of MIS, Eller College of Management, University of Arizona, Tucson AZ 85721