

# Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?

Hany M. SalahEldeen and Michael L. Nelson

Old Dominion University, Department of Computer Science  
Norfolk VA, 23529, USA  
{hany,mln}@cs.odu.edu

**Abstract.** Social media content has grown exponentially in the recent years and the role of social media has evolved from just narrating life events to actually shaping them. In this paper we explore how many resources shared in social media are still available on the live web or in public web archives. By analyzing six different event-centric datasets of resources shared in social media in the period from June 2009 to March 2012, we found about 11% lost and 20% archived after just a year and an average of 27% lost and 41% archived after two and a half years. Furthermore, we found a nearly linear relationship between time of sharing of the resource and the percentage lost, with a slightly less linear relationship between time of sharing and archiving coverage of the resource. From this model we conclude that after the first year of publishing, nearly 11% of shared resources will be lost and after that we will continue to lose 0.02% per day.

**Keywords:** Web Archiving, Social Media, Digital Preservation.

## 1 Introduction

With more than 845 million Facebook users at the end of 2011 [5] and over 140 million tweets sent daily in 2011 [16] users can take photos, videos, post their opinions, and report incidents as they happen. Many of the posts and tweets are about quotidian events and their preservation is debatable. However, some of the posts and events are about culturally important events whose preservation is less controversial. In this paper we shed light on the importance of archiving social media content about these events and estimate how much of this content is archived, still available, or lost with no possibility of recovery.

To emphasize the culturally important commentary and sharing, we collected data about six events in the time period of June 2009 to March 2012: the H1N1 virus outbreak, Michael Jackson's death, the Iranian elections and protests, Barack Obama's Nobel Peace Prize, the Egyptian revolution, and the Syrian uprising.

## 2 Related Work

To our knowledge, no prior study has analyzed the amount of shared resources in social media lost through time. There have been many studies analyzing the behavior of users within a social network, how they interact, and what content they share [3, 19, 20, 23]. As for Twitter, Kwak et al. [6] studied its nature and its topological characteristics and found a deviation from known characteristics of human social networks that were analyzed by Newman and Park [10]. Lee analyzed the reasons behind sharing news in social media and found that informativeness was the strongest motivation in predicting news sharing intention, followed by socializing and status seeking [4]. Also shared content in social media like Twitter move and diffuse relatively fast as stated by Yang et al. [22].

Further more, many concerns were raised about the persistence of shared resources and web content in general. Nelson and Allen studied the persistence of objects in a digital library and found that, with just over a year, 3% of the sample they collected have appeared to no longer be available [9]. Sanderson et al. analyzed the persistence and availability of web resources referenced from papers in scholarly repositories using Memento and found that 28% of these resources have been lost [14]. Memento [17] is a collection of HTTP extensions that enables uniform, inter-archive access. Ainsworth et al. [1] examined how much of the web is archived and found it ranges from 16% to 79%, depending on the starting seed URIs. McCown et al. examined the factors affecting reconstructing websites (using caches and archives) and found that PageRank, Age, and the number of hops from the top-level of the site were most influential [8].

## 3 Data Gathering

We compiled a list of URIs that were shared in social media and correspond to specific culturally important events. In this section we describe the data acquisition and sampling process we performed to extract six different datasets which will be tested and analyzed in the following sections.

### 3.1 Stanford SNAP Project Dataset

The Stanford Large Network Dataset is a collection of about 50 large network datasets having millions of nodes, edges and tuples. It was collected as a part of the Stanford Network Analysis Platform (SNAP) project [15]. It includes social networks, web graphs, road networks, Internet networks, citation networks, collaboration networks, and communication networks. For the purpose of our investigation, we selected their Twitter posts dataset. This dataset was collected from June 1st, 2009 to December 31st, 2009 and contains nearly 476 million tweets posted by nearly 17 million users. The dataset is estimated to cover 20%-30% of all posts published on Twitter during that time frame [21]. To select which events will be covered in this study, we examined CNN's 2009 events timeline<sup>1</sup>.

<sup>1</sup> <http://www.cnn.com/2009/US/12/16/year.timeline/index.html>