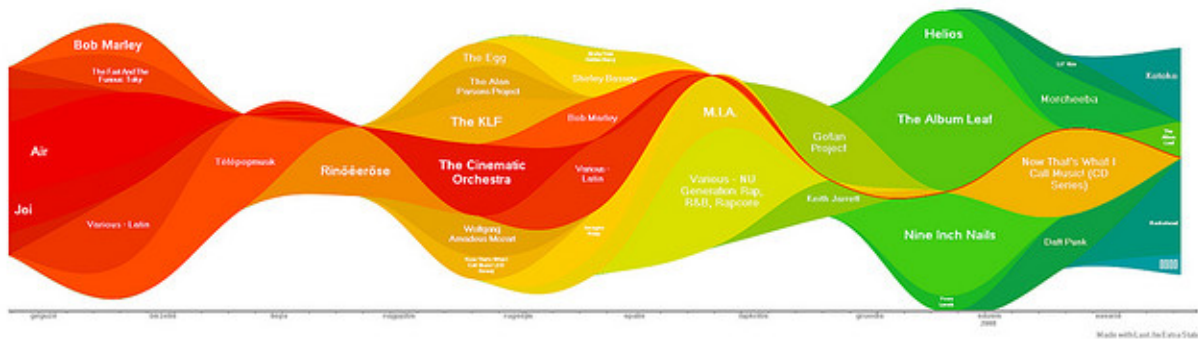# Leveraging the Internet's unique data repositories

July 10 2015, by Chris Givens

A Last.fm user shares the music recommendation system's representation of his listening habits over a month. Photo courtesy of Aldas Kirvaitis via Flickr.

Sometimes, data doesn't look like data. But when circumstances conspire and the right researchers come along, interesting facets of human nature reveal themselves. Last.fm and World of Warcraft are two entities made possible by the Internet, both aimed at entertainment of consumers. However, through new means of social interaction and larger scales of data collection they also, perhaps unintentionally, advanced science. Scientific achievement may seem like a stretch for a music service and a video game, but these unlikely candidates for scientific study show that the information age constantly offers new ways to study human behavior. Last.fm and World of Warcraft are contemporary social constructions, part of the new way that humans interact in our rapidly

changing digital world. By applying scientific rigor to the data unwittingly generated by two Internet-based companies, we see that knowledge is everywhere, but sometimes requires creative routes to coax it out of hiding.

Last.fm: more than a musical concierge

Last.fm is a [music service](#) that uses consumers' listening [data](#) and genre tags to recommend new music to the user. It has a huge cache of song clips in its databases, which were not viewed as a data set until recently, when a group of computer scientists mined the songs for certain characteristics and created a [phylogeny of popular music](#). The lead author on the study, Dr. Matthias Mauch, formerly worked on the Music Information Retrieval (MIR) team at Last.fm. MIR is essentially automated analysis of musical data, usually from audio samples. Uses for the data gleaned from audio samples include improved music search, organization, and recommendation. This kind of research has clear benefit to a company like Last.fm, whose main goal is to catalog users' listening habits and recommend music they would like based on past listening patterns. Dr. Mauch, however, is interested in more than simply improving musical recommendations; he wants to trace the evolution of the variety of music from around the world. In a [recent study](#), he used a huge data set obtained from his time at Last.fm to start cracking the code on musical evolution.

Hip-hop is a confirmed revolution

When hip-hop burst into the public consciousness in the late 1980's, the music polarized Americans. Hip-hop music originally centered on themes of social ills in inner-city America, providing a creative outlet for the frustration felt by many working-class African Americans at the time. Gangsta rap eventually grew out of hip-hop, characterized by at

times violent, masculine lyrical themes. After release of their seminal album, Straight Outta Compton, the hip-hop group N.W.A received a warning letter from the FBI as a result of controversial songs on the album. The explosive and politicized emergence of hip-hop created a new genre of popular music, thrusting a marginalized group of Americans into the pop culture spotlight. Starting from humble roots, hip-hop is now a multi-billion dollar industry. But even with all of the popular exposure and controversy, until Dr. Mauch's study the degree to which hip-hop revolutionized popular music was hard to quantify.

A group of researchers, led by Dr. Mauch, used MIR techniques on the Last.fm data set, and in doing so, found previously unknown relationships between hip-hop and other types of twentieth-century popular music. After recognizing the song clips obtained from Last.fm held a repository of data, the group devised a method of classifying songs based on two categories of attributes: harmonic and timbral. Harmonic attributes are quantifiable, encompassing chord changes and the melodic aspects of songs; timbral attributes are more subjective and focus on quality of sound, like bright vocals or aggressive guitar. The authors deemed these attributes "musically meaningful" and thus more appropriate for quantitative analysis than simple measures of loudness or tempo.

The researchers used modified text-mining techniques to carry out their analysis. They combined characteristics from the harmonic and timbral lists to create "topics" which could then be used to analyze each song based on the number of topics present. Next, the researchers analyzed 17,000 songs from the Billboard Hot 100 charts for the 50 years between 1960 and 2010. After finishing song analysis and clustering songs based on their harmonic and timbral characteristics, the researchers created a phylogenetic tree of popular music.

The tree empirically verified what we already knew—that hip-hop is in a

league of its own. Out of four clusters on the tree, hip-hop is the most divergent. Using the tree of life as an analogy, if the genres of rock, soul, and easy listening are animals, fungi, and plants, hip-hop would be musical bacteria.

Using these data, extensive knowledge of musical history is possible. The authors state in their paper that instead of using anecdote and conjecture to understand musical evolution, their methods make it possible to pinpoint precisely where musical revolutions occurred. Due to their efforts, popular music now has a quantitative evolutionary history, and Dr. Mauch isn't finished yet. He plans to do similar analyses on recordings of classical music and indigenous music from all over the world, in an attempt to trace the origins and spread of music predating the radio. I feel the innovative techniques and range of this study is incredible. Dr. Mauch and colleagues adapted research methods frequently used to improve of music delivery (already an interesting field) and used them to unlock a small amount of transcendent musical knowledge. This study shows that tens of thousands of song clips isn't a typical scientific data set, until someone says so. By taking what was provided and forging it into something workable, Dr. Mauch and colleagues applied scientific methods to Last.fm's unrecognized and unexamined data repository.

Surviving a pandemic in the Wide, World of Warcraft

World of Warcraft (WoW) is a highly social video game that connects players globally. WoW is also arguably the last place anyone would look for scientific insight. Launched in 2004, WoW is one of the most popular games ever created, with around ten million subscribers at its peak popularity. WoW was designed as a "massively multiplayer online role-playing game". When launched, players from all over the world began interacting in real time, throughout an intricately designed virtual world. The world was designed as a fantastical model of the real world,

complete with dense urban areas and remote, low population zones. In 2005, a glitch that caused a highly contagious sickness to be spread between players revealed this game to be an apt model of human behavior under pandemic conditions. The glitch drastically affecting gameplay for gamers and piqued the interest of several epidemiologists.

The "Corrupted Blood Incident"

The glitch came to be known as the "Corrupted Blood Incident" in the parlance of the game. It originated from one of the many things present in WoW that are not present in the real world: "dungeons". Dungeons in WoW are difficult areas populated by powerful "boss" characters that possess special abilities not normally found in the game. In 2005, one of these abilities, the "Corrupted Blood" spell, was modified by a glitch to have powers outside of the zone it normally resided in. Consequently, the highly contagious "Corrupted Blood" swept though WoW, killing many player characters and providing an accurate simulation of real-world pandemic conditions. "Corrupted Blood" infected player characters, pets, and non-player characters, which aided transmission throughout the virtual landscape. Only one boss character in one remote zone cast this spell, so its spread was a surprise to players and developers alike, adding to the accuracy of the "simulation".

The glitch stayed active for about a week, and during that time, gameplay changed dramatically. Because pets and non-player characters carried the disease without symptoms, reservoirs of the plague existed in the environment and helped nourish the outbreak. Players avoided cities for fear of contracting the disease. Some players who specialized in healing stayed in cities, helping especially weak players stay alive long enough to do business. Weaker, low-level players who wanted to lend a hand posted themselves outside of cities and towns, warning other players of the infection ahead. After a week of the pandemic in the game, the developers updated the code and reset their servers, "curing"

the WoW universe of this scourge of a glitch.

Some epidemiologists took note after observing the striking similarities between real world pandemics and the virtual pandemic in WoW. In the virtual pandemic, pets acted as an animal reservoir, as birds did in the case of avian flu. Additionally, air travel in WoW (which takes place on the back of griffins) proved analogous to air travel in the real world, thwarting efforts to quarantine those affected by the disease. Also, WoW is a social game full of tight-knit communities, and at the time had around 6.5 million subscribers, making it a reasonable virtual approximation of the social stratification that exist in real world society.

The behavior observed in WoW was not taken as a prescription for how to handle a pandemic or a prediction of what will happen. Rather, as Dr. Nina Fefferman put it in a 2010 TED talk, this event provided " inspiration about the sorts of things we should consider in the real world" when making epidemiological models. Dr. Fefferman's group discovered two behaviors displayed by players experiencing the virtual pandemic empathy and curiosity, which are not normally taken into account by epidemiological models. Curiosity was the most notable, because it paralleled the behavior of journalists in real world pandemics. Journalists rush into the infected site to report, and then rush out, hopefully before becoming infected, which is exactly what many players did in the infected virtual cities of WoW.

The "Corrupted Blood Incident" is the first known time that an unplanned virtual plague spread in a similar way to a real world plague. Though at first, most looked at this instance simply an annoying video game glitch. It took some creative scientists decided to see what they knowledge they could glean from the incident. Their observations suggest that sometimes, the best agent-based model is the one where actual people control the agents, and that simulations similar to computer games might "bridge the gap between real world epidemiological studies

and large scale computer simulations." Epidemiological models are now richer as a result of this knowledge. To learn more about how the "Corrupted Blood Incident" changed scientific modeling for pandemics, head on over to the PLOS Public Health Perspectives blog to hear Atif Kukaswadia's take on it.

Concluding thoughts

The Last.fm study and "Corrupted Blood Incident" show ways scientists can use esoteric corners of the Internet to illuminate interesting pieces of human history and behavior. New means of social interaction and new methods for collecting information bring about interesting, if slightly opaque, ways to discover new knowledge and advance scientific discovery. It is a credit that these scientists helped shed light on human history and interactions by looking past the traditional and finding data from novel sources.

This story is republished courtesy of PLOS Blogs: blogs.plos.org.

Provided by Public Library of Science