|  |  |  |
| --- | --- | --- |
| Team Name (optional):  Team Members**: Rachita Hajela, Nikita Keswaney, Anto Loyola**  Project Title: Hack Through Your Neighborhood  Project Description: The number of attempts at hacking into an internet network with respect to living expenditure of the locality and the internet usage of the average household in the locality. The locality is based on the zipcodes. Basically we are trying to find out if certain areas (with high living expenditure or high internet usage) are more prone to internet attacks than other areas.  Data Sources - Use the table below to list and describe potential data sources. | | |
| ***Part 2. General Data Source Information*** | | |
| **Data Sources** | **Data Source Description** | **Data Size** |
| Living Expenditures in Different Zip Codes in USA | There is a data source which provides the average living expenditure for the different regions (based on zip code) in USA  The data source consists of the following information related to what we want–  Population, Distribution of residents ages, home sales, taxes paid, household income distribution. We will either use the household income directly or will calculate the living expense on other parameters present in the data source as well. | ~ 1 TB |
| WINE – Worldwide Intelligence Network Environment | There is a data source given by Symantec which provides the various network threats and hack-attempts in various parts of the world – including USA  Symantec collects telemetry data from over 75 million machines. This data set records occurrences of all the known host- and network-based attacks, allowing researchers to map the spread of cyber threats around the world. For example, each record includes the signature of the attack, the OS version of the attack’s target, the name of the compromised process and the URL or file the attack came from. | ~ 4 TB |
| Internet Usage Data | There is a data source which provides the average internet usage per family in different areas of the world – including USA  We will be taking out the internet usage data based on the zipcodes of US. | ~ 4 TB |
| ***Part 3. Detailed Data Source Information*** | | |
| **Data Sources**  - From Part 2. above | **Data Characteristics**  - The Symantec source is somewhat real time.  - The rest are more like logs.  - It is statistically loaded data over a few years this millennium | **Data Frequency**  - Symantec updates its data source once a week ~ approximately |
| ***Part 4. Technologies*** | | |
| We will be Using MapReduce in Java as well as Pig for taking out the desired data in tabular form and may try to merge twitter as a data source but not sure about that yet. | | |
| ***Part 5. References*** | | |
| 1. Home Network Security - <http://www.cert.org/historical/tech_tips/home_networks.cfm>?  2. Building a Secure Home Network – http://www.sans.org/reading-room/whitepapers/hsoffice/building-secure-home-network-611  3. Studying Internet Use as Penetration Increases - <http://www.princeton.edu/~artspol/workpap/WP15%20-%20DiMaggio%2BHargittai.pdf>  4. Use of the internet in higher-income households - <http://www.pewinternet.org/2010/11/24/use-of-the-internet-in-higher-income-households/>  5. Cyberbullying - <http://www.pewinternet.org/2007/06/27/cyberbullying/>  6. Characterizing Transition Behaviors in Internet Attack Sequences (Rochester Institute of Technology, Rochester, New York 14623) PDF | | |