

RESEARCH REPORT

Vladimir Zadorozhny, Patrick Manning, Matthew Drwenski, and Evgeny Karataev

Towards a Social Weather Service:

Linking Social Sciences and Information Sciences in Large-Scale
Inequality Analysis

Introduction

The issues of inequality today propagate the dilemmas of the past and add new questions and difficulties. They require a complex and sophisticated analysis, ranging widely and using the best tools and the best minds of our age. Up until now, analysis of inequality has taken place within discrete disciplines. In studies of economic inequality, even those who study inequality by income and wealth tend to appear as separate camps. Yet at the level of popular conception, it is commonly assumed to be possible to link race, social status, education, physical condition, and income into overall assessments of equality and inequality. To make such an assessment may appear impossible at the scholarly level—given the number of variables, the quantity of data, and the diversity of measures. But the current

impetus toward interdisciplinary cooperation, backed by the advances in information-science methods and high performance computing (HPC), provide an opening toward resolution of this big issue in social science that should not be passed up.

Social dynamics driven by inequality may have a dramatic impact on development of human societies. As an example, consider a commonly observed inequality pattern in social development as it moves through periods of notable transitions. Figure 1a illustrates this pattern on a large-scale scenario of countries transiting from pre-industrial to post-industrial societies. This pattern has a form of inverted U-shape, indicating considerable inequality increase as societies move from a well-established way of life to new challenges and opportunities driven by advances in new technologies [2]. (Simon Kuznets, “Economic Growth and Income Inequality,” *American Economic Review* 45 (1955), 1-28) The inequality may have both positive and negative impact on a society. From one side, it may result in healthy competition, producing new ideas and increasing human wellbeing. However, as it approaches a certain threshold, the inequality may cause severe disruptions and instabilities, as illustrated in Figure 1b. The social instability may cause significant degradation of human wellbeing, involving civil unrest and slowing down social progress. Such instability patterns can occur at different scales and may vary in duration and severity.

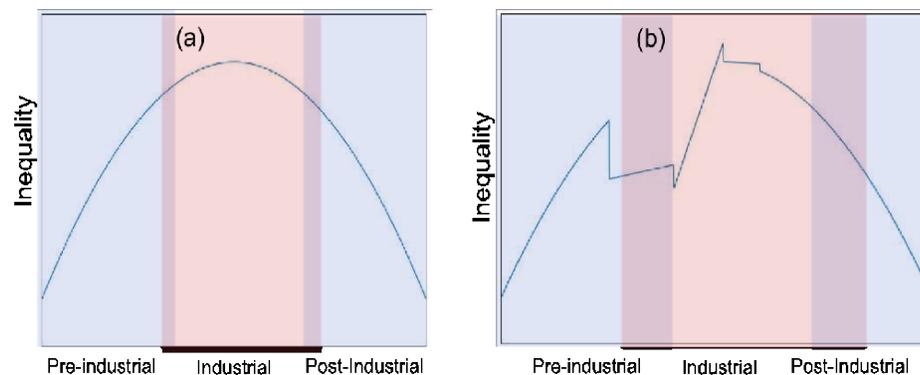


Figure 1: Social Dynamics during Transition Periods

The importance of timely discovery of the factors that may cause social instability is hard to underestimate. In this paper we address the challenge and outline the agenda of design and development of a social early warning and forecasting system, which we refer to as Social Weather Service (SWS) in the rest of this document.

Major Data Processing Challenges for the Social Weather Service

The value of the Social Weather Service depends on its efficiency in processing large amounts of social and historical data. Below we enumerate the key challenges in the socio-historical data processing that should be addressed in the SWS design:

1. **Distributed Data:** Socio-historical data may be spread over various data sources and organizations;
2. **Heterogeneous Data:** Socio-historical data may be represented in different data formats and in files with different types and structures;
3. **Sparse Data:** Socio-historical data may be fragmented—missing values are common;

4. **Aggregated Data:** Socio-historical data may be aggregated in different ways over various time intervals and space regions;
5. **Inconsistent Data:** Socio-historical data items reported by different data sources may contradict each other, resulting in data inconsistencies;
6. **Unreliable Data:** Socio-historical data sources may have different degrees of reliability.

We have addressed challenges (1) and (2) in the Tycho project integrating reports on diseases in the United States from 1888 to the present [5,4] and in the CHIA project funded by NSF, which resulted in a novel *Col*Fusion* infrastructure for historical data collection and integration [1,6]. In this project we will focus mostly on challenges (3)-(6), considering them in the context of inequality analysis. The inequality should be assessed based on combination of different factors (variables) from different categories. We propose subdividing inequality into eight major categories, as reflected in conceptual hierarchy in Figure 2: Economic Inequality (including earned and owned resources or income and wealth); Physical Inequality (including health variables and exposure to such issues as varying nutrition, expectation of life, and stature); and Social Inequality (including racial, status, gender, and civil condition). Each of these eight categories can be represented by one or more variables, and one can hope ultimately to combine and weight the various variables so as to develop an overall index of inequality.

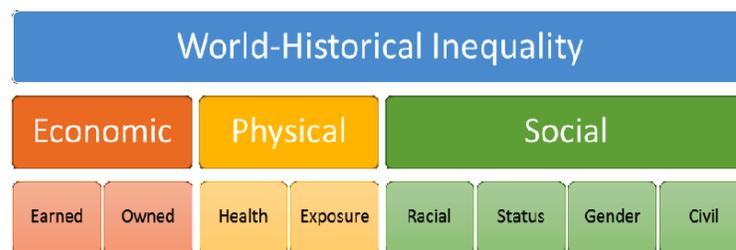


Figure 2: Conceptual Hierarchy of Inequality

In order to use the inequality-related variables for analysis of social stability, they should be measured comprehensively for different countries (or analogous units) over a sequence of time intervals (e.g., years). One key information-science breakthrough we seek is to link the currently discrete social science datasets into larger databases that can ultimately reach a global scale, through data-driven linkage. Conceptually, this task can be thought of as a challenge of creating an Integrated Table for Inequality Analysis (ITIA), in which available multidisciplinary data can be laid out to facilitate the data linkage.

The ITIA table should integrate available historical datasets, which are commonly small, complex, and, until recently, compiled by human agency. These historical data definitely qualify as big data when one considers very large quantities of data on the past that await retrieval and analysis [3]. So far we have identified over 210 separate published data sets related to the inequality variables from over 20 different sources. We have integrated those data sets and, looking at the national-level scale of analysis only, we created an ITIA table with over eight million data points.

It turned out that linking this information is problematic, even after the major data integration challenges (cleaning data, unifying data units and formats, accounting for territorial changes over time, etc.) are resolved. We

will need to combine incomplete data of dissimilar quality and scale, and of differing granularity. Drilling down to levels of greater detail and examining a single country, we observed the scattered nature of the data, which only grow sparser as we move further back in time. The task becomes even more complicated as we move beyond the national-level scale to examine smaller geographic units and to compare inequality among individuals across national boundaries at the global scale.

To address the *challenge of sparse data* we need to develop and apply advanced data imputation methods. Such methods can range from relatively simple techniques, such as using the sample mean, sliding window mean, or last observation for missing values, to more sophisticated approaches utilizing related variables, observations and semantic constraints based historical evidences. We should define proper metrics, reflecting similarities between countries/nations with respect to different inequality variables and time intervals. Such similarities can be refined iteratively starting from reasonable initial assumptions and providing more accurate estimates based on further analysis. We will also need large-scale simulation for approximation of the ITIA table. The simulation infrastructure will be used for imputation of missing data under most likely assumptions and application constraints. It will also allow us to assess the quality of various data imputation strategies using the complete simulated ITIA table as a “ground truth.”¹

To illustrate the *challenge of aggregated data*, consider a simple example in Figure 3, with average life expectancy statistics (LE) and Top 1% income share statistics (TIS) reported at different time intervals. We would like to estimate what value of LE better corresponds to given TIS. In database terms, we would need to join two tables together on their corresponding time intervals. However, performing a common equi-join on the reports’ start and end time would yield an empty result.

Life Expectancy			Total 1% Income Share		
From	To	LE	From	To	TIS
1981	1985	59.9	1985	1986	9.1
1986	1990	61.5			

Figure 3: Merging two variables at different time aggregation levels

We will need to develop advanced approximate join techniques that would intelligently provide the best effort to join tables on different aggregation levels. In general, to join aggregated data streams, we can either join aggregated reports directly (aggregated join), or first disaggregate reports to a common time unit and then use equi-join (disaggregated join). The process of merging aggregated data streams is resource consuming and it involves trade-offs between accuracy of the produced results, execution time, and consumed computational resources.

The *challenges of inconsistent and unreliable data* are tightly related. *Data inconsistency* is commonly caused by *inaccurate* historical reports, which may indicate a non-reliable source of data. In many cases, data inconsistency can be revealed through analysis of relationships between existing reports in the redundant database. Historical data sets may have different levels of *reliability* due to the quality of components such as the primary source of information, data collection methodology, etc. In order to assess the reliability of a report, we need to account for the data inconsistencies it causes. Assuming that the system continuously receives new historical reports, we can compute a reliability value for the source of these streams, which evolves with respect to new evidence. We need to explore advanced methods of information fusion and sense-making for *inconsistent* historical databases. Related tasks include (1) finding efficient strategies to check for inconsistencies in large-scale aggregated and sparse

historical databases, and (2) finding efficient inconsistency resolution strategy. A considerable challenge is to optimize the inconsistency and unreliability processing so that it can scale-up for large historical databases.

Scalable Infrastructure for Implementing Social Weather Service

As becomes obvious from considering the above challenges, efficient socio-historical data processing requires scalable high-performance computing infrastructure. Developing advanced infrastructure for conducting large-scale social-science analysis is crucial for success of this project. This infrastructure should efficiently combine computational resources, data repositories, and information processing methods from a variety of disciplines. The SWS infrastructure will considerably enhance Col*Fusion, supporting *research data analysis and computing at various scales* for efficient implementation of functionalities of the proposed Social Weather Service. It will also facilitate accessibility and usability of the SWS data and computational resources. It will implement easily extendable hardware and software configurations, and provide an on-demand access to data and computational resources reflecting specific needs of researchers. The SWS infrastructure will also provide an integrated data store for the large-scale data integration, sharing and processing, which is crucial for implementing SWS components. We plan to utilize the PSC's unique computational and data handling resources for the advanced information processing required for automatic large-scale data linkage, data reliability assessment, and data fusion. Such system and data integration service will efficiently support *collaborative interdisciplinary research data management*.

Comparison of geographic resources

As an initial follow-up to the workshop, Kathy Weimer (Rice University) and Tonia Sutherland (University of Alabama) conducted a preliminary survey of several geographic resources, to establish their relative strengths, especially in providing metadata to document place names. They gave special attention to spatial catalogues of the Library of Congress, Wikipedia, Geonames, and Getty. Their results indicated substantial overlap among the various resources—they cite each other in detail. All of them appear to be valuable but each has specific advantages over others. This investigation will continue in greater detail.

Initial list of world-historical places

As an initial list of geographic terms, David Ruvolo (University of Pittsburgh) completed a scan of all the index terms in a leading historical atlas, *Atlas of World History*, edited by Jeremy Black. The index, scanned through an Optical Character Recognition program and copied to a large Excel file, yielded 14,000 entries, of which over 11,000 are place names. This is proposed as the core list for the world-historical gazetteer. This data has been organized in four ways: (1) a single column of over 11,000 place names; (2) a single column of 14,000 entries including place names and other entries (battles, peoples, cultures); (3) multiple columns, in which place names are accompanied by other descriptive data, including ID labels. These additional data also include information on the time period relevant to each place (since page numbers are linked to atlas chapters organized by time period).

In steps to come, the list of place names can undergo a process of disambiguation, especially to clarify duplicate names. The team of workshop participants will then decide on next steps to take in constructing the world-historical gazetteer.

 Articles in this journal are licensed under a Creative Commons Attribution 4.0 United States License.



This journal is published by the University Library System of the University of Pittsburgh as part of its D-Scribe Digital Publishing Program and is cosponsored by the University of Pittsburgh Press.

¹ By “ground truth,” we mean that the simulated data are treated, for these purposes, as basic data.