Big Data Analyses for Collective Opinion Elicitation in Social Networks

Yingxu Wang
Dept. of Electrical and Computer Engineering
Schulich School of Engineering, Univ. of Calgary
Calgary, Alberta, Canada T2N 1N4
e-mail: yingxu@ucalgary.ca

Victor J. Wiebe
Dept. of Electrical and Computer Engineering
Schulich School of Engineering, Univ. of Calgary
Calgary, Alberta, Canada T2N 1N4
e-mail: victor_mx@shaw.ca

Abstract—Big data are extremely large-scaled data in terms of quantity, complexity, semantics, distribution, and processing costs in computer science, cognitive informatics, web-based computing, cloud computing, and computational intelligence. Censuses and elections are a typical paradigm of big data engineering in modern digital democracy and social networks. This paper analyzes the mechanisms of voting systems and collective opinions using big data analysis technologies. A set of numerical and fuzzy models for collective opinion analyses is presented for applications in social networks, online voting, and general elections. A fundamental insight on the collective opinion equilibrium is revealed among electoral distributions and in voting systems. Fuzzy analysis methods for collective opinions are rigorously developed and applied in poll data mining, collective opinion determination, and quantitative electoral data processing.

Keywords—Big data; big data engineering; numerical methods; fuzzy big data; social networks; voting; opinion poll; collective opinion; quantitative analyses

I. INTRODUCTION

Big data is one of the representative phenomena of the information era of human societies [8, 16]. Almost all fields and hierarchical levels of human activities generate exponentially increasing data, information, and knowledge. Therefore, big data engineering has become one of the fundamental approaches to embody the essences of the abstraction and induction principles in rational inferences where discrete data represent continuous mechanisms and semantics.

A field of big data applications is in human memory and DNA analyses in neuroinformatics, cognitive biology, and brain science, where huge amount of data and information have been obtained and pending for efficient processing [1, 3, 10, 17]. For instance, the biological information contained in a DNA is identified as up to 33 Peta-bit, i.e., 32,985,348,833,280,000 bit or 32,985,348 Giga-bit, of genetic information according to a formal neuroinformatics model [33].

Another paradigm of big data generated in computing is the Internet traffic as shown in Table I as of statistics in 2012 [30]. The big data over the Internet indicate human communication and information searching demands via digital devices such as over 4.6 billion mobile phones and equivalent number of tablets and portable computers. The big data in this domain has pushed the daily traffic from the rate of Terabyte ($10^{12}$) to that of Petabyte ($10^{15}$).

<table>
<thead>
<tr>
<th>Data hub</th>
<th>Data traffic</th>
<th>Rate/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYSE</td>
<td>1.0</td>
<td>Terabytes</td>
</tr>
<tr>
<td>Twitter</td>
<td>7.0</td>
<td>Terabytes</td>
</tr>
<tr>
<td>Facebook</td>
<td>10.0</td>
<td>Terabytes</td>
</tr>
<tr>
<td>Google</td>
<td>24.0</td>
<td>Pettabytes</td>
</tr>
<tr>
<td>Total Internet traffic</td>
<td>667.0</td>
<td>Exabytes ($10^{18}$)</td>
</tr>
</tbody>
</table>

TABLE I. THE BIG DATA TRAFFIC ON INTERNET IN 2012

Censuses and general elections are the traditional and typical domains that demand efficient big data analysis theories and methodologies beyond number counting [5, 13]. Among modern digital societies and social networks, popular opinion collection via online polls and voting systems becomes necessary for policy confirmation and general elections.

One of the central sociological principals adopted in popular elections and voting systems is the majority rule where each vote is treated with an equal weight [2, 4, 7]. The conventional methods for embodying majority rule may be divided into two categories known as the methods of max counting and average weighted sum. The former is the most widely used technology that determines the simple majority by the greatest number of votes on a certain opinion among multiple or binary options. The latter assigns various weights to optional opinions, which extends the binary selection to a wide range of weighted rating. Classic implementations of these voting methods are proposed by Borda, Condorcet, and others [5, 11, 12]. Borda introduced a scale-based system where each casted vote is attached a rank that represents an individual's preferences [5]. Condorcet developed a voting technology that determines the winner of an election as the individual who is paired against all alternatives as a run-off vote [11]. However, formal voting and general elections mainly adopt the mechanism that implements a selection of only-one-out-of-n options without any preassigned weight. In this practice for casting the majority rule in societies, the average weighted sum method is impractical.

This paper analyzes the formal mechanisms of voting systems and collective opinion elicitation in the big data engineering approach. The cognitive and computing properties of big data in general, and of the electoral big data
in particular, are explored in Section II. A set of mathematical models and numerical algorithms for collective opinion analyses is developed in Section III and illustrated in Section IV. Fuzzy models for collective opinion elicitation and aggregation are rigorously described in Section V. A set of real-world case studies on applications of the formal methodologies is demonstrated in big poll data mining, collective opinion determination, and quantitative electoral data processing.

II. PROPERTIES OF DATA IN BIG DATA ENGINEERING

This section explores the intentions and extensions of big data as a term. The sources of big data generation are analyzed. Special properties of big data are elaborated in computer science, cognitive informatics, web-based computing, and computational intelligence.

A. The Computational Properties of Big Data

Definition 1. Data, D, are an abstract representation of the quantity Q of real-world entities or mental objects by a quantification mapping \( f_q : Q \rightarrow D \) (1)

Although decimal numbers and systems are mainly adopted in human civilization, the basic unit of data is a bit [9, 15], which forms the converged foundation of computer and information sciences. Therefore, the most fundamental form of information that can be represented and processed is binary data. Based on bit, complex data representations can be aggregate to higher structures such as byte, natural numbers (\( \mathbb{N} \)), real numbers (\( \mathbb{R} \)), structured data, and databases.

The physical model of data and data storage in computing and the IT industry are the container metaphor where each bit of data requires a bit of physical memory.

Definition 2. Big data are extremely large-scaled data across all aspects of data properties such as quantity, complexity, semantics, distribution, and processing costs.

The basic properties of big data are unstructured, heterogeneous, monotonous growing, mostly nonverbal, and decay in information consistency or increase of entropy over time [20]. The inherent complexity and exponentially increasing demands create unprecedented problems in all aspects of big data engineering such as big data representation, acquisition, storage, searching, retrieve, distribution, standardization, consistency, and security.

The sources of big data are human collective intelligence. Typical mathematical and computing activities that generate big data are Cartesian products (\( O(n^m) \)), sorting (\( O(n \log n) \)), searching (exhaustive, \( O(n^2) \)), knowledge base update (\( O(n^2) \)), as well as permutation and NP problems with \( O(2^n) \), \( O(n!) \), or even higher orders [9]. Typical human activities that produce big data are such as many-to-many communications, massive downloads of data replications, digital image collections, and networked opinion forming.

Although the syntax of data is concrete based on computation and type theories, the semantics of data is fuzzy [24, 25, 27, 32, 33]. The analysis and interpretation of big data may easily exceed the capacity of conventional counting and statistics technologies.

B. The Cognitive Properties of Big Data

The neurophysiological metaphor of data as factual information and knowledge in human memory is a relational network [10, 17, 19, 26], which can be represented by the Object-Attribute-Relation (OAR) model [19, 29] as shown in Figure 1.

Figure 1. The OAR model of data and knowledge in memory

Definition 3. The OAR model of data and knowledge as retained in long-term memory in the brain is a triple, i.e.:

\[ OAR \triangleq (O, A, R) \] (2)

where \( O \) is a finite set of objects denoting the extension of a data concept, \( A \) is a finite set of attributes for characterizing the data concept, and \( R \) is a set of relations between the objects and attributes.

The OAR model enables the estimation of the memory capacity of human, which revealed the nature of big data as cognitive and semantics entities in the brain. In cognitive neurology, it is observed that there are about \( 10^{11} \) neurons in the brain, each of them is with 10^3 synaptic connections in average [3, 10]. According to the OAR model, the estimation of the capacity of human memory for big data representation can be reduced to a classical combinatorial problem as follows.

Definition 4. The capacity of human memory \( C_m \) is determined by the total potential relational combinations, \( C_m^r \), among all neurons \( n = 10^{11} \) and their average synaptic connections \( s = 10^3 \) to various related subset of entire neurons, i.e.:

\[ C_m = C_m^r = \frac{10^{41}}{10^8(10^{11} - 10^3)!} = 10^{4.432} \text{ [bit]} \] (3)

Eq. 3 provides an analytic explanation of the upper limit of the potential number of synaptic connections among neurons in the brain. The model reveals that the brain does not create new neurons to represent new information;
in order to represent the newly acquired information.

Both cognitive and computational foundations of data explored in this section explain the nature of big data and the need for big data engineering. The notion of big data engineering is perceived as a field that studies the properties, theories, and methodologies of big data as well as efficient technologies for big data representation, organization, manipulations, and applications in industries and everyday life. It is noteworthy that, although the appearance of data is discrete, the semantics and mechanisms behind them are mainly continuous. This is the essence of the abstraction and induction principles of natural intelligence.

### III. METHODS FOR BIG ELECTORAL DATA ANALYSES

Mathematical models and numerical methods for rigorous voting data processing and representation are sought in this section in order to reveal the nature of big data in voting and collective opinions. This leads to a set of novel methods beyond traditional counting technologies such as regressions of opinion spectrums, adaptive integrations of collective opinions, and allocation of the opinion equilibrium.

#### A. Big Data Interpretation for Embodying the Majority Rule in Sociology

As reviewed in Section I, the typical method for implementing the majority rule via voting is used to be the max finding method.

**Definition 5.** The max function elicits the greatest number of votes on a certain opinion, $O_i$, $1 \leq i \leq n$, as the voting result $V^m$ among a set of $n$ options, i.e.:  

$$V^m = \max(N_{O_1}, N_{O_2}, ..., N_{O_n})$$  

where $N_{O_i}$ is the number of votes casted for opinion $O_i$.

When there are only two options for the voting, $V^m$ is significantly greater than $O_0$ according to Eq. 5. Although the maximum vote appears at 0 over the opinion spectrum, the real representative centroid of the collective opinion is actually at about 1.3 on the spectrum. In other words, the mean of the entire votes indicated an equilibrium point of the collective opinion in between those of $O_0$ and $O_1$.

**Definition 6.** The pseudo majority dilemma states that the result of a voting based on the simple max mechanism may not represent the majority opinion distribution casted in the voting, i.e.:  

$$V^m = \max(N_{O_0}, N_{O_1}, N_{O_2}, N_{O_3}, N_{O_4})$$

$$= N_{O_i, \text{max}} \sum_{i=1}^{n} N_{O_i} | i \neq i_{\text{max}},$$

where $\sum_{i=1}^{n} N_{O_i} - N_{O_{i, \text{max}}} \geq N_{O_{i, \text{max}}}$

A typical case of the pseudo majority dilemma in voting can be elaborated in the following example.  

**Example 1.** A voting with a distributed political spectrum from far right ($N_{O_0}$), right ($N_{O_1}$), neutral ($N_{O_2}$), left ($N_{O_3}$), and far left ($N_{O_4}$) is shown in Figure 2 where the vote distribution is $X = [N_{O_0}, N_{O_1}, N_{O_2}, N_{O_3}, N_{O_4}] = [4000, 2500, 2600, 1200, 1100]$. According to the max finding method given in Eq. 4, the voting result is:

$$V^m = \max(N_{O_0}, N_{O_1}, N_{O_2}, N_{O_3}, N_{O_4})$$

$$= \max(4000, 2500, 2600, 1200, 1100)$$

$$\Rightarrow O_0 (N_{O_0} = 4000)$$

The result indicates that opinion $O_0$ is the winner and the other votes would be ignored. However, in fact, the sum of the rest opinions $\sum_{i \neq 0}^{n} N_{O_i} = 2500 + 2600 + 1200 + 1100 = 7300$, is significantly greater than $O_0$ according to Eq. 5. Although the maximum vote appears at 0 over the opinion spectrum, the real representative centroid of the collective opinion is actually at about 1.3 on the spectrum. In other words, the mean of the entire votes indicated an equilibrium point of the collective opinion in between those of $O_0$ and $O_1$.

Therefore, in order to rationally analyze popular opinion distributions and the representative collective opinion on an opinion spectrum, advanced mathematical models, numerical methods, and fuzzy analyses [6, 21, 32, 33] are yet to be rigorously studied for voting data processing and representation.

#### B. Numerical Regression for Analyzing Opinion Spectrum Distributions beyond Counting

On the basis of analyses in the preceding subsection, an overall perspective on the collective opinions casted in a voting can be rigorously modeled as a nonlinear function over the opinion spectrum. In order to implement a complex polynomial regression, a numerical algorithm is developed in MATLAB as shown in Figure 3, which can be applied to analyze any popular opinion distribution against a certain political spectrum represented by bid voting data. In the
analysis program, a 3rd order polynomial is adopted for curve fitting, while other orders may be chosen when it is appropriate. The general rule is that the order of the polynomial regression $m$ must less than the points of the collected data $n$. Data interpolation technologies may be adopted to improve the smoothness or missing points of raw data in numerical technologies [6, 28, 32].

The seats distribution of Canadian parties in the House of Commons is given in Table II [30]. In Table II, the relative position of each party on the political spectrum is obtained based on statistics of historical data such as their manifesto, policy, and common public perspectives [14, 18].

According to the data in Table II, i.e., $X = [-100, -71, -43, -14, 0, 50]$ and $Y = [100, 4, 1, 34, 4, 160]$, the voting results can be rigorously represented by the following function, $f(x)$, as a result of the polynomial regression implemented in Figure 3.

$$f(x) = 0.0001x^3 + 0.005x^2 + 2.175x + 65.1182 \quad (6)$$

where $m = 3$ and $n = 5$.

The above regression analysis results are visually plotted in Figure 4. Because the polynomial characteristic function is a continuous characteristic function, it can be easily processed for multiple applications such as for opinion spectrum representation, equilibrium determination, and analyses of policy gains based on the equilibrium benchmark as described in the following subsection.

![Figure 4. House seats of parties on the political spectrum of Canada](image)

C. The Collective Opinion Equilibrium Elicited from a Spectrum of Opinion Distributions

It is recognized that the representative collective opinion on a spectrum of opinion distributions casted in an election is not a simple average of weighted sum as conventionally perceived. Instead, it is the centroid covered by the curve of the characteristic regression function as marked by the red $\oplus$ sign as shown in Figure 4.

**Definition 7.** The opinion equilibrium $\Xi$ is the natural centroid in a given weighed opinion distribution where the total votes of the left and right wings reached a balance at the point $k$, i.e.:

$$\Xi(x \mid x = k) \triangleq \int_{x=-n}^{k} \nu(x)dx \int_{x=k}^{n} \nu(x)dx , x, k \in [-n, n] \quad (7)$$

where

$$\int_{x=-n}^{k} \nu(x)dx = \frac{1}{2} \int_{x=-n}^{n} \nu(x)dx = \frac{1}{2} (\int_{x=-n}^{k} \nu(x)dx + \int_{x=k}^{n} \nu(x)dx)$$

The integration of distributed opinions based on the regression function can be carried out using any numerical integration technologies. For instance, the iterative Simpson’s integration method for an arbitrary continuous function $f(x)$ over $[a, b]$ can be described as follows:

$$I_n = \int_{a}^{b} f(x)dx, k = \frac{n-1}{3}, \quad I^b = 0 \quad (8)$$

where

$$R_n f = I_{1}^{b} + \frac{h}{3} (f(x_n)) + 4f(x_{n}) + f(x_{n}))$$

The collective opinion equilibrium method as modeled in Eq. 7 is implemented in the algorithm as shown in Figure 5. The core integration method adopted in the algorithm is based on a built-in function `quad()` in MATLAB [6] that implements Eq. 8.

**Example 3.** Applying the collective opinion equilibrium determination algorithm to the opinion distribution data in the Canadian general election as given in Figure 4, the
opinion equilibrium is obtained as $\Xi_j = 20.3$. The result indicates that the overall national opinion equilibrium was at the mid-right as casted in 2011.

```matlab
function [TotalOpinionIntegration, Equilibrium] = vote_equilibrium_analysis(xl, xu)
    % The integration of total opinion in the voting
    TotalOpinionIntegration = quad(f, xl, xu);  % Simpson iterative integration
    % To find the opinion equilibrium by iterative integration
    h = 0.1;
    for MidPoint = xl : h : xu
        IGi = quad(f, xl, MidPoint);  % Simpson iterative integration
        if IGi >= TotalOpinionIntegration / 2
            break
        end
    end
    Equilibrium = MidPoint;
end
```

Figure 5. Algorithm of collective opinion equilibrium determination

Because the collective opinion equilibrium $\Xi$ is the centroid of the opinion integration as defined in Eq. 7, it is obvious that the equilibrium cannot be simply determined or empirical allocated without the computational algorithm (Figure 5) as demonstrated in Example 3.

IV. ELECTORAL DATA PROCESSING

Using the methodologies developed in Section III, useful applications will be demonstrated in this section with real-world data. The case studies encompass the analysis of a series of general elections in order to find out the dynamic equilibrium shifts and the extrapolation of potential policy gains based on the historical electoral data.

A. Analysis of a Series of Historical Elections Based on Equilibrium Benchmarking

A benchmark of opinion equilibrium can be established on the basis of a series of the historical electoral data. Based on it, trends of the political equilibriums can be rigorously analyzed in order to explain: a) What was the extent of serial shifts as casted in the general elections? and b) Which party was closer to the political equilibrium represented by the collective opinions casted in the general elections?

Example 4. The trend in Canadian popular votes over time can be benchmarked by results from the last four general elections as given in Table III. Applying the opinion equilibrium determination algorithm, $vote_{equilibrium\_analysis}$ as given in Figure 5, the collective opinions distributed in Figure 6 can be rigorously elicited, which indicates a dynamic shifting pattern of the collective opinion equilibriums, i.e., $5.0 \rightarrow 7.4 \rightarrow 7.0 \rightarrow 10.9$, on the political spectrum between [-100, 100] during 2004 to 2011.

The opinion equilibrium determination method provides insight for revealing the implied trends and the entire collective opinions distributed on the political spectrum. An interesting finding in Example 4 is, although several parties on the left spectrum, -100 ≤ x < 0, had won significant number of votes as shown in Table III and Figure 6, the collective opinion equilibrium had mainly remained unchanged at the area of central-right where $\Xi = 7.6$ in average.

<table>
<thead>
<tr>
<th>Political Party</th>
<th>Votes 2004</th>
<th>Votes 2006</th>
<th>Votes 2008</th>
<th>Votes 2011</th>
<th>Relative position on the spectrum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conservatives</td>
<td>4,019,498</td>
<td>5,374,071</td>
<td>5,209,069</td>
<td>5,832,401</td>
<td>-50</td>
</tr>
<tr>
<td>Liberals</td>
<td>4,982,220</td>
<td>4,479,415</td>
<td>3,633,185</td>
<td>2,783,175</td>
<td>-14</td>
</tr>
<tr>
<td>Green</td>
<td>582,247</td>
<td>664,068</td>
<td>937,613</td>
<td>576,221</td>
<td>-43</td>
</tr>
<tr>
<td>Bloc Quebecois</td>
<td>1,680,109</td>
<td>1,553,201</td>
<td>1,379,991</td>
<td>889,788</td>
<td>-71</td>
</tr>
<tr>
<td>New Democrats</td>
<td>2,127,403</td>
<td>2,589,597</td>
<td>2,515,288</td>
<td>4,508,474</td>
<td>-100</td>
</tr>
<tr>
<td>Opinion equilibrium</td>
<td>5.0</td>
<td>7.4</td>
<td>7.0</td>
<td>10.9</td>
<td></td>
</tr>
</tbody>
</table>

TABLE III. HISTORICAL ELECTORAL DATA DISTRIBUTIONS OF CANADIAN GENERAL ELECTIONS

B. Extrapolation for Potential Policy Gains Based on Benchmarked Collective Opinion Equilibrium

The key objective of a party in a general election is to rationally predict what the potential gain would be for a certain policy making or shifting. The theory of the collective opinion equilibrium as developed in preceding sections suggests that this target can be reached by adapting current policies towards the equilibrium benchmark.

Definition 8. A target gain in elections can be extrapolatively projected via a necessary shift of policy $\Delta x = n' - n$ that satisfies the equilibrium benchmark $\Xi$ by an updated regression of expected opinion distributions $\nu_a(x)$:

- a) At the right end: $\Delta x = n' - n$, when $\int_{x=k}^{n'} \nu_a(x) dx = I_o / 2, \ n' \in [k, n]$
- b) At the left end: $\Delta x = n' - n$, when $\int_{n}^{x=k} \nu_a(x) dx = I_o / 2, \ n' \in [-n, k]$
- c) In between: $\Delta x = a' - a$, and $a, a' \in (-n, n)$

where the original position $n$, point of equilibrium $\Xi (k | x=k)$, and the total integrated votes $I_o$ are known as results of analyses in Section III.

Figure 6. Polynomial regressions for federal elections during 2004 to 2011

634
The extrapolation method is divided into three cases according to the position of the party on the spectrum, which can be at either end or in the middle of the spectrum.

**Example 5.** Given a target of a 10% gain in terms of number of votes in the future election for a party where \( n = 50 \) on the spectrum, what kind of policy manipulations may be needed to contribute towards the expected objective based on the equilibrium benchmark casted in the latest election as obtained in Example 4?

Based on the historical data provided in Table III, i.e., \( X = [-100, -71, -43, -14, 50], Y_{2011} = [4508474, 889788, 576221, 2783175, 5832401] \), and the total opinion integration \( I_n = 441854649 \), the projected electoral improvement problem can be represented as follows:

\[
\begin{align*}
X^\prime &= [-100, 50, 71, 43, 14] \\
Y^\prime &= [4508474, 889788, 576221, 2783175, 5832401 \times 1.1 = 6415641] \\
I^\prime &= 441854649
\end{align*}
\]

Solve the problem according to Eq. (9a), the following results are obtained: \( n' = 48.2 \) and \( \Delta X = n' - n = 48.2 - 50 = -1.8 \). That is, in order to gain 10% more votes, the party would need to shift its policy leaning to the collective opinion equilibrium \( \Xi = 10.9 \) for 1.8 steps where the negative sign indicates a move to the middle. In case where other factors would change as well, the problem becomes a multi-party gaming system. However, for any given moment, the system is still determinable based on the same analysis method and algorithm as presented in Sections III and IV.

### V. FUZZY METHODS FOR COLLECTIVE OPINION ELICITATION AND ANALYSIS BASED ON BIG POLL DATA

Big data analysis technologies for collective opinion elicitation based on historical data have been demonstrated in preceding sections, which reveal that a party may gain more votes by adapting its policy towards the political equilibrium established in past elections. It is recognized that a social system is conservative which is not change rapidly over time because the huge base of population and human cognitive tendency according to the long-life span system theory [23]. However, the collective opinion equilibriums do shift dynamically. Therefore, an advanced technology for enhancing potential policy gains is to calibrate the current collective opinion equilibrium by polls in order to support up-to-date analysis and predication.

#### A. Fuzzy Elicitation of Collective Opinion from Big Poll Data Samples

The typical technology for detecting current collective opinion equilibrium is by polls. A poll may be designed to test the impact of a potential policy in order to establish a newly projected equilibrium. The projected equilibrium will be used to update and adjust the historical benchmark. In this approach, rational predictions of policy gains towards a general election or a social network voting can be obtained in a series of analytic regressions as formally described in the remainder of this subsection.

**Definition 9.** An opinion \( o_i \) on a given policy \( p_i \) is a fuzzy set of degrees of weights \( \omega^i_k \) expressed by \( j, 1 \leq j \leq m \), groups in the uniform scale \( I \), i.e.:

\[
o_i \triangleq f : p_i \rightarrow I, I = [0,1]
\]

\[
= \{(p_i, \omega^i_1), (p_i, \omega^i_2), \ldots, (p_i, \omega^i_m)\}
\]

\[
= \{R(p_i, \tilde{o}_i)\}
\]

where the big-\( R \) notation represents recurring entities or repetitive functions indexed by the subscript [20].

The normalized scale for fuzzy analyses is a universal one because any other scale can be mapped into it.

**Definition 10.** A collective opinion \( \tilde{O}_p \) on a set of \( n \) policies \( p_i, 1 \leq i \leq n \), is a compound opinion as a fuzzy set of average weights \( \tilde{o}^i_j = \frac{1}{n} \sum_{k=1}^{n} \tilde{o}(i, j) \) on each policy, i.e.:

\[
\tilde{O}_p = \{R_{i=1}^{n} R_{j=1}^{m} \tilde{o}_i^j\}
\]

\[
= \{R_{i=1}^{n} \} \frac{1}{n} \sum_{k=1}^{n} \tilde{o}(i, j), \tilde{o}_i^j \in [0,1]
\]

\[
= \{(p_1, \tilde{o}_1^1), (p_1, \tilde{o}_1^2), \ldots, (p_1, \tilde{o}_1^m)\}
\]

\[
= \{(p_2, \tilde{o}_2^1), (p_2, \tilde{o}_2^2), \ldots, (p_2, \tilde{o}_2^m)\}
\]

\[
= \ldots
\]

\[
= \{(p_n, \tilde{o}_n^1), (p_n, \tilde{o}_n^2), \ldots, (p_n, \tilde{o}_n^m)\}
\]

where \( \tilde{O}_p \) may be aggregated against the averages of each row or column that indicate the collective opinions of a certain policy casted by all groups or that of all policies of a certain group, respectively, as illustrated in Table IV.

**Definition 11.** The effect \( \tilde{E} \) of a set of policies is a fuzzy matrix of the average weighted differences between the current opinion \( \tilde{w}_{ij} \) and the historical ones \( \tilde{w}_{ij}^* \) for the \( i \)th policy on the collective opinion of the \( j \)th group, i.e.:

\[
\tilde{E} \triangleq \tilde{O}_i - \tilde{O}_i^*, 1 \leq i \leq n, 1 \leq j \leq m
\]

\[
= \{R_{i=1}^{n} R_{j=1}^{m} \tilde{o}_i^j - \tilde{o}_i^j\}
\]

**Definition 12.** The impact \( \tilde{I} \) of a policy is a fuzzy matrix of products of effects \( \tilde{E} \) and the corresponding group sizes \( N_j \), i.e.:

\[
\tilde{I} \triangleq \{R_{i=1}^{n} R_{j=1}^{m} N_j \tilde{E}_j\}
\]

\[
= \{R_{i=1}^{n} R_{j=1}^{m} N_j (\tilde{w}_j - \tilde{w}_{ij})\}
\]

where the ± sign indicates a positive or negative impact on a target group, respectively.

**Definition 13.** The gain of policy impacts, \( \tilde{G} \), is a fuzzy
set of the mathematical means of the cumulative impacts that each group obtain as results of the series of aggregations from the initial poll data, i.e.:
\[
\tilde{G} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{n} \sum_{j=1}^{n} I_{ij}
\]  
(14)

B. Fuzzy Analyses of Potential Policy Impacts in Votes

The fuzzy methodologies for collective opinion elicitation and analysis from big poll data as developed in Section V.A are illustrated in application case studies in the following examples.

Example 6. The collective opinion \(\tilde{O}_p\) on the set of 3 testing policies against 5 groups on the political spectrum can be elicited based on a set of large sample poll data as summarized in Table IV. The current average weights of opinions \(\tilde{\omega}^\text{ij}\) and those of the historical ones \(\tilde{\omega}^\text{ij}\) are aggregated from the sample data of individual opinions according to Eqs. 10 and 11.

<table>
<thead>
<tr>
<th>(\tilde{\omega}^\text{ij})</th>
<th>(G_1)</th>
<th>(G_2)</th>
<th>(G_3)</th>
<th>(G_4)</th>
<th>(G_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p_1)</td>
<td>0.2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.4</td>
<td>0.9</td>
</tr>
<tr>
<td>(p_2)</td>
<td>1.0</td>
<td>0.9</td>
<td>0.5</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>(p_3)</td>
<td>0.5</td>
<td>0.5</td>
<td>0.6</td>
<td>0.6</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**Definition 14.** The complexity or size of poll data, \(C_p\), is proportional to the numbers of testing policies \(|P|\), groups on the spectrum \(|G|\), and number of sample individuals \(N_p\), i.e.:
\[
C_p = |P| \cdot |G| \cdot N_p
\]  
(15)

where 2,000 tests in a poll will result in 30,000 raw individual opinions in the settings of Example 6.

Example 7. Based on the summarized poll data as given in Table IV with the average collective opinions, the fuzzy set of effects \(\tilde{E}\) of the \(i\)th policy to the collective opinion of the \(j\)th group can be quantitatively determined according to Eq. 12 as follows:
\[
\tilde{E} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{n} \sum_{j=1}^{n} (p_{ij} \cdot \tilde{\omega}^\text{ij} - \tilde{\omega}^\text{ij}'), \ n = 3, m = 5
\]
\[
= \left\{ \begin{array}{l}
(0.2, 0.4, 0.6, 0.7, 0.5) \\
(1.0, 0.5, 0.7, 0.5, 0.6) \\
(0.5, 0.6, 0.3, 0.7, 1.0) \\
(0.1, 0.2, 0.1, -0.1, 0.1) \\
(0.3, 0.3, -0.2, 0.1, 0.2)
\end{array} \right.
\]
where the most effective policy is \(p_3 \rightarrow \{G_1, G_2\}\) with a 30% improvement, while the most negatively effective policy is \(p_1 \rightarrow G_3\) with a -40% loss.

Example 8. On the basis of Table IV and Example 7, the impact \(I\) of each tested policy is a fuzzy matrix of the products of individual group size and the effects that projects the \(i\)th policy on the \(j\)th group with the size \(N_{gj}\), i.e.:
\[
\tilde{I} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{n} \sum_{j=1}^{n} (p_{ij} \cdot \tilde{E}_{ij}, n = 3, m = 5)
\]
\[
= \left\{ \begin{array}{l}
(0.1 \times 450847, 0.07 \times 489788, 0.2 \times 576221, 0.1 \times 2783175, -0.4 \times 5832401) \\
(0.1 \times 450847, -0.2 \times 489788, -0.1 \times 576221, 0.1 \times 2783175, 0.1 \times 5832401) \\
(0.3 \times 450847, 0.3 \times 489788, -0.2 \times 576221, 0.1 \times 2783175, 0.2 \times 5832401) \\
450847, 0, 115244, 278318, -2332960 \\
450847, -177958, -57622, 278318, 5832401 \\
135242, 300060, -115244, 278318, 1166480
\end{array} \right.
\]

where \(N_{gj} = [4508474, 889788, 576221, 2783175, 832401]\) according to the 2011 data in Table III.

Example 9. The potential average gain of policy impacts \(G\) can be derived according to Eq. 14 based on the results in Example 8 as follows:
\[
\tilde{G} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{n} \sum_{j=1}^{n} \tilde{I}_{ij}, n = 3, m = 5
\]
\[
= \left\{ \begin{array}{l}
450847, 0, 115244, 278318, -2332960 \\
135242, 300060, -115244, 278318, 1166480
\end{array} \right.
\]

where \(N_{gj} = [4508474, 889788, 576221, 2783175, 832401]\)

The projected gains or losses, \(G\), over the political spectrum produce a new set of estimated electoral distributions \(Y = Y + G = [4508474, 889788, 576221, 2783175, 5832401] + [751412, 296600, -19207, 278317, -194413] = [5259886, 919488, 557014, 3061492, 5637988]\).
On the basis of the projected gains derived from current polls of collective opinions, the potential shift of the collective opinion equilibrium on the political spectrum can be predicted using the algorithm in Figure 5. The regression result is plotted in Figure 7, which indicates a collective opinion equilibrium shift slightly to the middle, i.e., $\Delta E = E_2 - E_1 = 9.8 - 10.9 = -1.1$, by contrasting to that of the historical vote distributions.

VI. CONCLUSIONS

Big data engineering has been introduced into the field of sociology for collective opinion elicitation and analyses. Numerical models and fuzzy methodologies have been developed for rigorously analyzing voting and electoral data. This approach has led to the revealing of deep implications, complex equilibrium, and dynamic trends represented by popular opinion distributions on a political spectrum. A key finding in this work has been the existence of the collective opinion equilibrium over a spectrum of opinion distribution in big poll data, which is not simply a weighted average rather than the point of natural centroid at the integrated areas of opinion distributions. Adaptive policy gains based on historical and current poll data have been formally derived from fuzzy collective opinion aggregation, effect analyses, and quantitative impact estimations. A set of interesting insights has been demonstrated on the nature of large-scale collective opinions in poll data mining, collective opinion equilibrium determination, and quantitative electoral data processing in big data engineering.

ACKNOWLEDGMENT

This work is supported in part from a discovery fund granted by the Natural Sciences and Engineering Research Council of Canada (NSERC). We would like to thank the anonymous reviewers for their valuable suggestions and comments on the previous version of this paper.

REFERENCES