INVITED PAPER

Words on the Web: Noninvasive Detection of Emotional Contagion in Online Social Networks

This paper reviews and discusses a nonexperimental, noninvasive method to detect and quantify contagion of semantic expression in massive online social networks.

By Lorenzo Coviello, *Student Member IEEE*, James H. Fowler, and
Massimo Franceschetti, *Senior Member IEEE*

ABSTRACT | Does semantic expression spread online from person to person? And if so, what kinds of expression are most likely to spread? To address these questions, we developed a nonexperimental, noninvasive method to detect and quantify contagion of semantic expression in massive online social networks, which we review and discuss here. Using only observational data, the method avoids performing emotional experiments on users of online social networks, a research practice that recently became an object of criticism and concern. Our model combines geographic aggregation and instrumental variables regression to measure the effect of an exogenous variable on an individual's expression and the influence of this change on the expression of others to whom that individual is socially connected. In a previous work, we applied our method to the emotional content of posts generated by a large sample of users over a period of three years. Those results suggest that each post expressing a positive or negative emotion can cause friends to generate one to two additional posts expressing the same emotion, and it also inhibits their use of the opposite emotion. Here, we generalize our method so it can be applied to contexts different than emotional expression and to different forms of content generated by the users of online platforms. The method allows us to determine the usage of words in the same semantic category spread, and to estimate a signed relationship between different semantic categories, showing that an increase in the usage of one category alters the usage of another category in one's social contacts. Finally, it also allows us to estimate the total cumulative effect that a person has on all of her social contacts.

KEYWORDS | Influence; instrumental variables; nonexperimental methods; semantic expression; social networks

I. INTRODUCTION

In the last decade, the challenge of understanding the spreading and synchrony of human behavior over social networks has attracted the attention of the research community at large. The problem originally arises in the context of the social sciences, but due to the expanding usage of online social networks, it has also attracted the interest of the engineering community with the aim of quantifying these effects using the massive amount of data that these networks generate. Studies have included the diffusion of news and "memes" [1]; cascades in communication platforms, networked games, microblogging services [2]; health-related phenomena such as obesity and smoking [3], [4]; emotional states like happiness and depression [5], [6]; purchase of online products [7], [8]; clicking online advertisements; and joining online recreational leagues and store purchases [9].

Manuscript received March 2, 2014; revised October 24, 2014; accepted October 28, 2014. Date of publication November 13, 2014; date of current version November 18, 2014. This work was supported in part by the Army Research Office (ARO) under Grant W911NF-11-1-0363.

L. Coviello and **M. Franceschetti** are with the Department of Electrical and Computer Engineering, University of California San Diego, San Diego, CA 92092 USA (e-mail: lcoviell@ucsd.edu).
J. H. Fowler is with the Department of Political Science and School of Medicine, Medical Genetics Division, San Diego, CA 92093 USA.

Digital Object Identifier: 10.1109/JPROC.2014.2366052

Studies based on observational data pose an inherent difficulty for causal inference because social contacts may have similar behavior as a result of at least two processes: homophily (the tendency of similar individuals to group together) or influence [10], [11]. Controlled experiments allow us to disentangle influence effects from homophily both in the laboratory [12] and online [13]–[16], but they are often limited in scale and lack external validity. Large scale experiments have been shown to be feasible in the context of political participation [16], product adoption [7], [8] and emotional influence [17], but are often impractical or require very close collaboration with private companies.

Moreover, the experimental change in the users' experience required by some of these studies recently came under scrutiny because of questions about the ethics involved. Some people criticized [18] a large scale study [17] of emotional contagion on Facebook in which the researchers changed the content shown to some users in order to study their reaction. Similar criticisms were directed at the online dating website OkCupid for experimenting with their platform in order to understand how individuals react to each other [19]. These recent events call for the development of alternative, nonexperimental methods to study human behavior at large scale [20].

Our work in [21] was an attempt to compensate for the shortcomings of existing experimental and observational approaches, using a method to detect and quantify influence via instrumental variable regression. We studied text-based expression in massive social networks, developed a model of emotional contagion of semantic expression, and validated it on the content posted by a large sample of Facebook users over a period of three years.

In this paper, we show how our model can also be applied to different and possibly heterogeneous data from other social networking platforms, and to contexts other than emotional expression. Our approach is fully nonexperimental: it is based only on observational data and, as a result, it does not alter users' experience. It also guarantees respect for user privacy: for our study in [21], individuals' information and posts were never visible to researchers and resided on secure servers where Facebook stores user data, and were analyzed only at an aggregate level. The study was reviewed for ethics and approved in advance by the Institutional Review Board at the University of California San Diego, San Diego, CA, USA.

Focusing on the mathematical model and on the engineering methodology employed, this paper reviews and complements our previous work. Our individual-level model assumes that a person's usage of words in a semantic category is a linear function of temporal and individual baseline effects; exogenous variables like news, the stock market, or the weather; and endogenous variables—corresponding to the usage of given semantic categories in posts written by the person's social contacts, referred to as "friends." The reciprocal causality between the endogenous variables of the model makes it difficult to obtain consistent and unbiased estimates of social influence. Therefore, we proceed in two steps. First, we aggregate the model on a geographical basis by averaging over all people who are in the same city, obtaining a model based on the same coefficients as the individual-level model but with a much smaller number of observations. Second, we deal with the problem of reciprocal causality by estimating the model using instrumental variable regression, a method pioneered in economics [28]. This method relies on the availability of an exogenous variable—called an instrument—that affects the endogenous variables (friends' posts) but does not directly induce a change in the subject's posts, called the dependent variable. In general, valid instruments might be unavailable, or they might lack sufficient power to predict changes in the endogenous variable. In our work, we considered rainfall experienced by friends as the instrument, using data made available by the National Climatic Data Center (NCDC),[1] which proved to be a robust predictor of emotional expression. Upon finding a relationship between friend's rainfall and their expression, we can assume the former affects the latter as the opposite direction is unlikely. Our method first computes the effect that friends' rainfall (the instrument) has on friends' posts (the endogenous variables). Then, it evaluates the corresponding effect of the rainfall-induced change in friends' expression on the person's posts (the dependent variable).

In order to obtain consistent estimates, the instrument must satisfy the exclusion restriction [28]. This posits that, controlling for all other variables, the instrument (friends' rainfall) must not directly affect the dependent variable. An implication of this restriction is that the instrument must also be uncorrelated with the exogenous variable experienced by the subject (subject's rainfall), otherwise the model might only be estimating how a subject's rainfall affects her own expression. Therefore, to break any correlation between a subject's rainfall and friends' rainfall, we restricted our analysis to observations for which it did not rain in the subject's city. Once this is applied, the subject's rainfall is constant in the data set and, therefore, it does not correlate with friends' rainfall. Moreover, breaking the correlation between user's and friends' rainfall solves the potential issue of the geographic similarity of the weather in close-by cities. As a result, we must also focus exclusively on social ties between individuals in different cities (see Section III-D). Note that individuals in different cities likely do not interact face to face, but they can reach each other via multiple communication media, such as the telephone, e-mail, and social networking websites. Therefore, any influence detected between them is unlikely to be caused by physical

[1] http://www.ncdc.noaa.gov

interaction and would suggest that remote communication plays an important role in spreading semantic expression.

Our method allows us to determine what semantic categories are susceptible to influence between social contacts by estimating how an individual's usage of a semantic category is affected by her friends' usage of the same category. We can then use the estimates for each semantic category to rank them from the most to the least likely to spread.

Moreover, our method allows us to determine the relationship between different semantic categories, by estimating how an individual's usage of one category is altered by her friends' usage of a different category. This will help us to understand whether the usage of a semantic category fosters or inhibits the usage of other categories. We already showed in [21] that expression of positive affect inhibits expression of negative affect and *vice versa*.

Finally, our model allows us to compute the cumulative effect a person has on her friends (see Section III-F). Although the effect on any one social contact will be small, each person typically has many social contacts, so the total expected effect of a single act of expression may alter the expression of several other people. Here, we show how to use our model to quantify this multiplier effect on posts within the same semantic category and on posts in different categories.

A. Related Work

Our work is related to a growing body of literature on influence and diffusion in networks, whose goal is to characterize how behaviors and information spread from person to person. Online social networks are becoming increasingly popular as research environments and sources of data for these investigations. For example, the content posted by people online has been used to identify which people or topics are influential in social networking websites [29] and in the blogosphere [30]. It has also been used to study which network attributes and sharing behaviors make people influential [31], which topics (e.g., represented by hasthtags) diffuse in a more persistent way [32], and even to study the structure of diffusion cascades on different communication platforms [2]. Large scale experimental studies have isolated the role of the network in the diffusion of information [33], emotional expression [17], and behaviors [7], [16]. However, homophily has been shown to play a similarly important role, and scholars have devoted their attention to distinguishing between the two phenomena and to comparing the size of their effects [11], [14], [34], [35].

Our work is related to the econometric literature on instrumental variables. Instrumental variables have been proposed as a tool to infer causal effects from observational data [28]. This approach has been applied to a variety of contexts, such as labor economics [36], the study of the causal effect of education on earning [37], program evaluation [38], the characterization of neighborhood effects [39], and the impact of microfinance [40]. However, valid instruments can be difficult to find [41], and scholars have warned against the risks of using "weak" instruments that do not predict variation in the endogenous variable with sufficient precision [42].

A large body of research studies text meaning by analyzing patterns of words or grammar [43]–[45]. However, the performance of most traditional classification methods relies on sufficient text length, as in the case of bag of words or kernel-based methods [46], [47]. The analysis of short text from microblogging services (such as Twitter or Facebook) requires new approaches [48]–[50], which in some cases leverage metadata (e.g., user's information) or the content of related posts.

Although we mainly focus on the engineering aspects of the detection and measurement of peer influence in semantic expression, our work is also related to sociolinguistics. The full understanding of language in a society requires us to consider the social network in which the language is embedded, intended as the set of relationships and interactions between its individuals [51]. Scholars have argued that speech patterns might depend on the looseness and tightness of the social network [52]. Our model formulation allows us to take tie strength between individuals into account. Different approaches have been proposed to quantify tie strength in online social networks [53], [54], and future research should investigate whether strong ties play a major role in the spread of semantic expression.

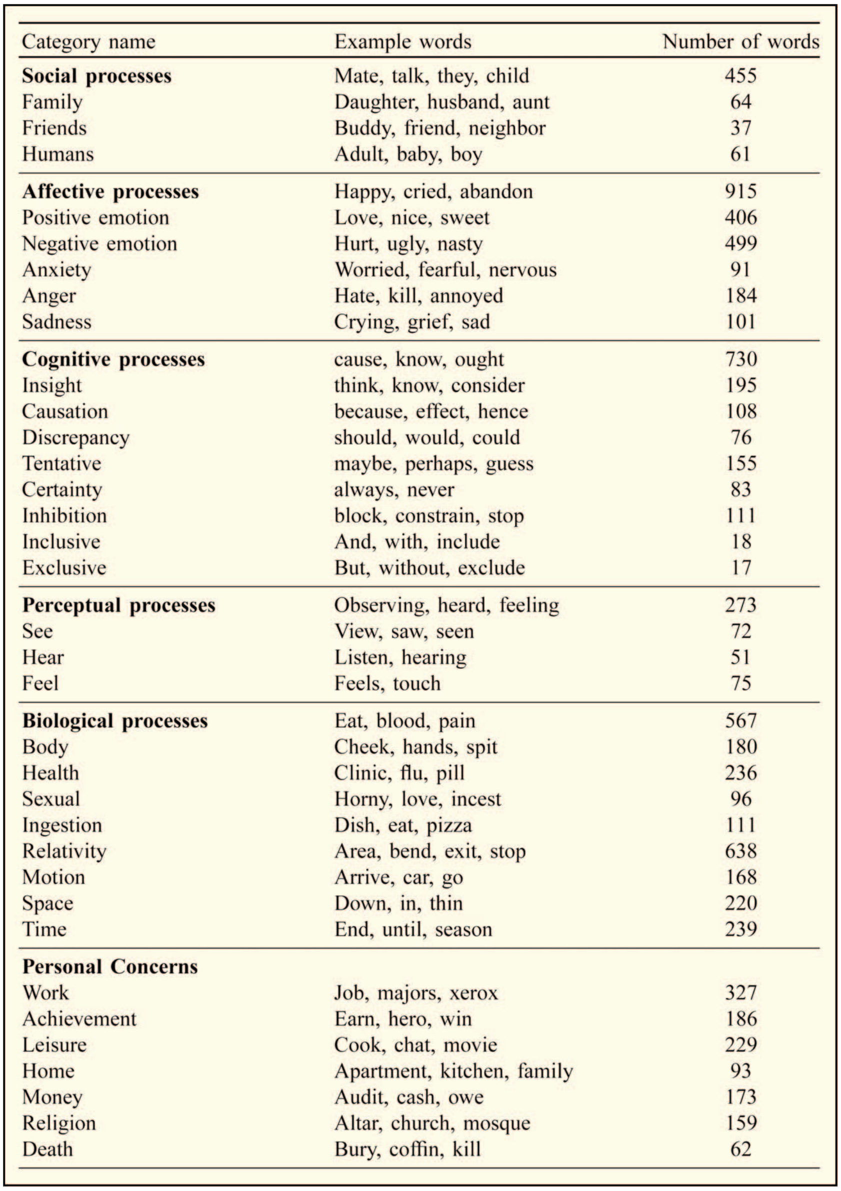
II. MODEL VARIABLES

We consider a set T of distinct days. For each day $t \in T$, let $S(t)$ be the population on day t, and let $n(t) = |S(t)|$ be their number. To apply our method, we assume that individuals can be geolocated at the level of cities. For each city g let $S_g(t)$ be the set of individuals in city g on day t and let $n_g(t) = |S_g(t)|$. In general, one might consider different time and geographic resolution. We assume resolution at the level of days and cities in accordance to our previous work [21].

A. Quantifying the Semantic of Text-Based Expression

Several methods can be used to quantify semantic expression of the content posted by individuals (see discussion in Section V). We referred to the semantic categories defined by the Linguistic Inquiry and Word Count (LIWC) 2007 [22], a word classification tool widely used in the social sciences and in psychology research [23]–[27]. The LIWC contains several classes of processes, each of which contains one or more semantic categories, pertaining to affective processes, perceptual processes, biological processes, social processes, and personal concerns. A list of semantic categories from the LIWC is given in Table 1. In [21], we considered the categories for positive and negative affective processes. In general, a

Table 1 List of Semantic Categories From the LIWC



Category name	Example words	Number of words
Social processes	Mate, talk, they, child	455
Family	Daughter, husband, aunt	64
Friends	Buddy, friend, neighbor	37
Humans	Adult, baby, boy	61
Affective processes	Happy, cried, abandon	915
Positive emotion	Love, nice, sweet	406
Negative emotion	Hurt, ugly, nasty	499
Anxiety	Worried, fearful, nervous	91
Anger	Hate, kill, annoyed	184
Sadness	Crying, grief, sad	101
Cognitive processes	cause, know, ought	730
Insight	think, know, consider	195
Causation	because, effect, hence	108
Discrepancy	should, would, could	76
Tentative	maybe, perhaps, guess	155
Certainty	always, never	83
Inhibition	block, constrain, stop	111
Inclusive	And, with, include	18
Exclusive	But, without, exclude	17
Perceptual processes	Observing, heard, feeling	273
See	View, saw, seen	72
Hear	Listen, hearing	51
Feel	Feels, touch	75
Biological processes	Eat, blood, pain	567
Body	Cheek, hands, spit	180
Health	Clinic, flu, pill	236
Sexual	Horny, love, incest	96
Ingestion	Dish, eat, pizza	111
Relativity	Area, bend, exit, stop	638
Motion	Arrive, car, go	168
Space	Down, in, thin	220
Time	End, until, season	239
Personal Concerns		
Work	Job, majors, xerox	327
Achievement	Earn, hero, win	186
Leisure	Cook, chat, movie	229
Home	Apartment, kitchen, family	93
Money	Audit, cash, owe	173
Religion	Altar, church, mosque	159
Death	Bury, coffin, kill	62

larger set C of semantic categories can be considered by our method.

For day $t \in T$ and subject $i \in S(t)$, let $U_i(t)$ be the set of all content posted by subject i on day t, and let $u_i(t) = |U_i(t)|$ be its cardinality. For each subject i such that $u_i(t) > 0$, and each category $c \in C$, let $u_i^{(c)}(t)$ be the number of elements of $U_i(t)$ containing at least one word from category c, and let

$$y_i^{(c)}(t) = \frac{u_i^{(c)}(t)}{u_i(t)}$$

be the frequency, or usage, of category c by subject i on day t. Note that $0 \leq y_i^{(c)}(t) \leq 1$. Therefore, a subject i such that $u_i(t) > 0$ is characterized by $|C|$ variables $y_i^{(c)}(t)$ quantifying her usage of words form all categories in C during day t. Observe that a single piece of user content can contribute to the frequency $y_i^{(c)}(t)$ for several categories c.

B. Exogenous Control Variable

Our method relies on the availability of an exogenous variable that affects the semantic expression of a person's friends but not (directly) the semantic expression of the person. We call this variable the "instrument." Our model characterizes how a change in the instrument induces a change in friends' semantic expression, and how the induced change predicts a change in the person's semantic expression.

There are many sources of exogenous variation in the world, but we chose rainfall as the instrument, relying on data from NCDC. For each city g, we consider the NCDC station closest to it, and let $\bar{x}_g(t) = 1$ if that station recorded rainfall on day t, and zero otherwise. For each subject $i \in S_g(t)$, let $x_i(t) = \bar{x}_g(t)$, that is, a binary indicator variable of rainfall in city g. We focus on rainfall as the instrument for several reasons. First, its geographical resolution lends itself to the analysis of our geographically aggregated model. Second, individuals in the same city tend to experience the same weather on a given day. Moreover, in [21], we show it is a robust instrument in the sense that it captures enough variation of the endogenous explanatory variable (friends' emotional expression). Other meteorological variables would have been a valid alternative. The identification of valid instruments is challenging and finding a systematic way to characterize them is key to apply our method to more general contexts.

C. Social Network Information

For each day $t \in T$, and subjects $i, j \in S(t)$, let $a_{i,j}(t) \in [0,1]$ be the strength of the relationship from i to j on day t, which need not be symmetric. Also, let $\delta_i(t) = \sum_{j \in S(t)} a_{i,j}(t)$. In [21], we let $a_{i,j}(t) \in \{0,1\}$, where $a_{i,j}(t) = 1$ denotes that i and j were friends on day t. In this case, $\delta_i(t)$ is the degree of subject i on day t (that is, the total number of friends of the subject). Allowing $a_{i,j}(t)$ to have any value between zero, one would allow to asses the role of tie strength.

III. MODEL

A. Individual-Level Model

Recall that $y_i^{(c)}(t)$ represents the usage of category c by subject i on day t. We assume that $y_i^{(c)}(t)$ is a function of several terms, according to

$$\begin{aligned} y_i^{(c)}(t) = {} & \theta(t) + f_i + \beta_{c',c} x_i(t) \\ & + \gamma_{c',c} \frac{1}{\delta_i(t)} \sum_j a_{i,j}(t) y_j^{(c')}(t) + \epsilon_i(t). \end{aligned} \quad (1)$$

$\theta(t)$ represents a "fixed effect" for day t and takes into account temporal patterns of variation in the use of category c (for example, people might be more likely to write about work during the weekdays, or more likely to write about health in the winter). f_i represents a fixed effect for subject i and takes into account different baseline usage of category c for different people (for example, some people might write about work more than others). $x_i(t)$ represents the exogenous variables experienced by subject i on day t. Equation (1) assumes that the effect of the exogenous variable $x_i(t)$ is weighted by a coefficient $\beta_{c',c}$ (the same for all subjects i), whose sign and strength represent the effect of the exogenous variable on usage of semantic category c. The summation in (1) represents the effect of usage of semantic category $c' \in C$ by i's friends on i's usage of category c.[2] Note that the effect of friends' expression is assumed to be inversely proportional to i's degree $\delta_i(t)$, compatible with the idea that a person with a lot of friends is less likely to view posts by all of them. This endogenous term is weighted by the coefficient $\gamma_{c',c}$, which represents the direction and strength of influence (assumed to be the same for all subjects). Finally, $\epsilon_i(t)$ are assumed independent and identically distributed normal error terms with zero mean and variance σ^2, to take unobserved factors into account.

The main parameter of interest is the coefficient $\gamma_{c',c}$ for all $c, c' \in C$, which expresses how a change in the semantic expression of i's friends affects subject i's semantic expression. However, the reciprocal causality present in model (1) makes it difficult to obtain unbiased estimates of the model parameters. This is due to the inherent feedback present in the model. That is, there is mutual influence between any pair of subjects i and j, and influence might follow even more complex paths (for example, i's expression in category c might influence j's expression in category c', which might affect k's expression in category c''). We address this issue in two steps, by first proposing an aggregated version of model (1) that averages over people who are in the same city (see Section III-B), and then by relying to the method of instrumental variable regression [28] (see Section III-C).

We also observe that model (1) is memoryless. This is a simplifying assumption that makes the method of instrumental variable regression easily applicable. Moreover, the model has one observation for each subject i and day $t \in T$, which, given a set of hundreds millions of users, would be difficult to analyze without some form of aggregation.

B. Geographical Aggregation

We average (1) over all $n_g(t)$ subjects $i \in S_g(t)$ who are in city g on day t

$$\begin{aligned} \frac{1}{n_g(t)} \sum_{i \in S_g(t)} y_i^{(c)}(t) = {} & \frac{1}{n_g(t)} \\ & \times \sum_{j \in S_g(t)} \left(\theta(t) + f_i + \beta x_i(t) + \frac{\gamma}{\delta_i(t)} \sum_j a_{i,j}(t) y_i^{(c')}(t) + \epsilon_i(t) \right). \end{aligned}$$

This can be written as

$$\bar{y}_g^{(c)}(t) = \theta(t) + \bar{f}_g + \beta \bar{x}_g(t) + \gamma \bar{Y}_g^{(c')}(t) + \bar{\epsilon}_g(t) \quad (2)$$

where we substituted $\gamma_{c',c}$ with γ and $\beta_{c',c}$ with β for ease of notation. In (2), $\bar{y}_g^{(c)}(t)$ is the average usage of category c by

[2] The model specification in (1) is not restricted to $c = c'$. It allows us to study the effect of usage of a semantic category c' on a potentially different category c.

subjects who are in city g; $\bar{f}_g$ is the average baseline usage of category c; $\bar{x}_g(t)$ is the average exogenous variable experienced by the subjects (rainfall in city g, in [21]); $\bar{\epsilon}_g(t)$ is the sum of $n_g(t)$ independent normal random variables with zero mean and variance σ^2 and, therefore, has variance $\sigma^2/n_g(t)$. The term $\bar{Y}_{gt}$ represents how usage of category c by subjects in city g is affected by the usage of category c' by their friends, and can be written as

$$\begin{aligned}\bar{Y}_g^{(c')}(t) &= \sum_j \frac{y_j^{(c')}(t)}{n_g(t)} \sum_{i \in S_g(t)} \frac{a_{i,j}(t)}{\delta_i(t)}\\ &= \sum_j y_j^{(c')}(t) A_{j,g}(t)\end{aligned}$$

where $A_{j,g}(t)$ represents the strength of the relationship from subject j to city g (normalized by the number of those subjects), that is, the influence from j to city g.

The coefficients β and γ are the same in (1) and (2). That is, the coefficients of the individual level model (1) can be estimated from the aggregated model (2). And note that our approach is unlikely to create an "ecological fallacy," which occurs when there are opposing effects at the individual and aggregated level, as individuals in the same city are very likely to experience the same weather [55]. Different instruments might lead to different situations.

Finally, the aggregated model (2) has a single observation for each city g and day t, a much smaller figure than the individual-level model (1), which would have millions of observations for each day in a large data set. This makes estimation more practical.

C. Instrumental Variable Regression

We are interested in estimating the parameter γ in (2). However, the explanatory variable $\bar{Y}_g^{(c')}(t)$ is an endogenous variable, that is, it can be correlated with both the dependent variable $\bar{y}_g^{(c)}(t)$ and the error term $\bar{\epsilon}_g(t)$. Since ordinary least squares regression would not produce unbiased estimates for γ, we use the method of instrumental variable regression [28]. This method can produce consistent and unbiased estimates even when there is reciprocal causation (as in our case, where people affect their friends and *vice versa*). All that is needed is an instrument that predicts the endogenous variable but not the dependent variable. More formally, given a linear model of the form

$$y = \alpha x + \lambda v + \epsilon$$

where v is an endogenous variable correlated with both the dependent variable y and the error term ϵ, an instrument for v is an exogenous variable z that does not appear in the model equation, is correlated with v (conditional on all the exogenous explanatory variables), and is not correlated with the error term [28]. Moreover, we look for a variable z such that, upon finding a relationship between z and v, z affects v and not *vice versa*. Once such variable z is available, instrumental variable regression estimates the original model in two stages. First, the endogenous variable v is projected onto the subspace of all exogenous explanatory variables, according to the model

$$v = \alpha_1 x + \alpha_2 z + \nu$$

where ν is an error term uncorrelated with any regressor. Then, the predicted values $\hat{v}$ resulting from the projection are used to estimate the model

$$y = \alpha x + \lambda \hat{v} + \epsilon.$$

In our model, an instrument for the endogenous explanatory variable $\bar{Y}_g^{(c')}(t)$ is an exogenous variable z that is uncorrelated with the error term in (2) [that is, $\mathrm{Cov}(z, \hat{\epsilon}_g(t)) = 0$] and is partially correlated with $\bar{Y}_g^{(c')}(t)$ when controlling for the other exogenous explanatory variables. In the context of our model, we can write

$$\bar{Y}_g^{(c')}(t) = \theta'(t) + \bar{f}'_g + \beta_2 \bar{x}_g(t) + \beta_1 z + \nu_g(t) \qquad (3)$$

where $\nu_g(t)$ is an error term that is uncorrelated with any regressors and $\theta'(t)$ and $\bar{f}'_g$ are time and subpopulation fixed effects.

Equation (3) can be seen as the linear projection of $\bar{Y}_g(t)$ on the space of all the exogenous variables. Substituting (3) into (2) yields

$$\begin{aligned}\bar{y}_g^{(c)}(t) = (\theta(t) + \gamma\theta'(t)) &+ \left(\bar{f}_g + \gamma \bar{f}'_g\right)\\ &+ (\beta + \gamma\beta_2)\bar{x}_g(t) + \gamma\beta_1 z + \bar{\epsilon}'_g(t) \qquad (4)\end{aligned}$$

where the error term is uncorrelated with all the explanatory variables.

As the instrument z for $\bar{Y}_g(t)$, we define a variable $\bar{X}_g(t)$ that combines rainfall experienced by the friends of subjects in city g

$$\begin{aligned}\bar{X}_g(t) &= \sum_j x_j(t) \frac{1}{n_g(t)} \sum_{i \in S_g(t)} \frac{1}{\delta_i(t)} a_{i,j}(t)\\ &= \sum_j x_j(t) A_{j,g}(t) = \sum_h \bar{x}_h(t) \sum_{j \in S_h(t)} A_{j,g}(t)\\ &= \sum_h \bar{x}_h(t) B_{h,g}(t)\end{aligned}$$

where the sum is over all cities h, and

$$B_{h,g}(t) = \frac{1}{n_g(t)} \sum_{i\in S_g(t)} \frac{1}{\delta_i(t)} \sum_{j\in S_h(t)} a_{i,j}(t)$$

represents the strength of the relationship from city h to city g. We use $\bar{X}_g(t)$ to predict $\bar{Y}_g(t)$. $\bar{X}_g(t)$ is uncorrelated with the error term in (2), and it is partially correlated with $\bar{Y}_g^{(c')}(t)$.

The procedure above is equivalent to estimating the model in (2) using two-stage least squares (2SLS) regression. The first-stage regression estimates a model of the form

$$\bar{Y}_g^{(c')}(t) = \theta'(t) + \bar{f}'_g + \beta_1\bar{X}_g(t) + \beta_2\bar{x}_g(t) + \epsilon'_g(t). \quad (5)$$

The second-stage regression uses the predicted values $\bar{Y}_g^{(c',\mathrm{pred})}(t)$ from the first stage to estimate the model

$$\bar{y}_g^{(c)}(t) = \theta(t) + \bar{f}_g + \beta\bar{x}_g(t) + \gamma\bar{Y}_g^{(c',\mathrm{pred})}(t) + \bar{\epsilon}_g(t). \quad (6)$$

Finally, recall that the variance of the error term $\bar{\epsilon}_g(t)$ is proportional to $1/n_g(t)$ where $n_g(t)$ is the number of individuals in a city. Therefore, we weight each observation by the corresponding value of $n_g(t)$. To conduct the analysis, we use the function ivreg2 written for STATA [56].

D. Dealing With the Exclusion Restriction

A key assumption of instrumental variables regression is the exclusion restriction [28], according to which the instrument $\bar{X}_g(t)$ must not directly influence the dependent variable $\bar{y}_g^{(c)}(t)$. In our case, a person and some of her friends are experiencing similar $\bar{x}_g(t)$ as they are in the same city or in close-by cities. Therefore, in order to break the correlation between $\bar{X}_g(t)$ and $\bar{x}_g(t)$, we only consider observations for city-day pairs (g,t) such that $\bar{x}_g(t) = 0$ (in [21], it did not rain in city g on day t). Conditional on $\bar{x}_g(t) = 0$, (5) and (6) can be written as

$$\begin{aligned}\bar{Y}_g^{(c')}(t) &= \theta'(t) + \bar{f}'_g + \beta_1\bar{X}_g(t) + \epsilon'_g(t) && (7)\\ \bar{y}_g^{(c)}(t) &= \theta(t) + \bar{f}_g + \gamma\bar{Y}_g^{(c',\mathrm{pred})}(t) + \bar{\epsilon}_g(t). && (8)\end{aligned}$$

Note that since $\bar{x}_g(t) = 0$ the instrument $\bar{X}_g(t)$ now depends only on friends who are in different cities (not in city g). Therefore, our approach can only detect and measure influence between individuals in different cities.

E. Robustness of the Instrument

In order to assess the quality of the estimates obtained via instrumental variable regression, we also compute diagnostic statistics. First, we need to verify that the model is not underidentified. We use the Kleinbergen–Paap *rk* LM statistic to test the null hypothesis of underidentification [57]. Second, we need to verify that the instruments are good predictors of the endogenous explanatory variable in the first-stage regression (otherwise the instruments are considered weak). Weak instruments would cause poor predicted values in the first-stage regression and therefore poor estimation in the second-stage regression. To ensure the instruments are not weak, the Cragg–Donald Wald F statistic must exceed the critical threshold suggested by Stock and Yogo [58].

F. The Effect of a Person on Her Friends

We show that the coefficient γ represents the expected total effect of a person on her friends. In other words, it is the number of additional posts containing a word in category c posted by all of j's friends on day t caused by subject j's own post. Recall the individual-level model (1)

$$\begin{aligned}y_i^{(c)}(t) = \theta(t) + f_i + \beta x_i(t)&\\ + \gamma\frac{1}{\delta_i(t)}\sum_j a_{i,j}(t)y_j^{(c')}(t) + \epsilon_i(t).& \quad (9)\end{aligned}$$

Letting j be a subject who writes a post on day t, we compare the cases in which j's post contains a word in category c' $(y_j^{(c')}(t) = 1)$ and that in which it does not $(y_j^{(c')}(t) = 0)$. Simple manipulation of (9) shows that this difference is given by $\gamma a_{i,j}(t)/\delta_i(t)$. Summing over all subjects i who wrote a post on day t, the total effect of $y_j^{(c')}(t) = 1$ for a given subject j is

$$E_j(t) = \frac{\gamma\sum_i a_{i,j}(t)}{\delta_i(t)}. \quad (10)$$

The expected total effect of a person on all her friends is obtained by averaging (10) over all subjects j

$$\begin{aligned}\bar{E}(t) &= \frac{1}{n(t)}\sum_j E_j(t) = \gamma\frac{1}{n(t)}\sum_j\sum_i a_{i,j}(t)/\delta_i(t)\\ &= \gamma\frac{1}{n(t)}\sum_i\frac{1}{\delta_i(t)}\sum_j a_{i,j}(t) = \gamma\frac{1}{n(t)}\sum_i\frac{\delta_i(t)}{\delta_i(t)} = \gamma.\end{aligned}$$

Therefore, we can refer to the coefficient γ as the expected total effect of a person on her friends.

IV. RESULTS

In this section, we review the results from [21] to show how the method works. In future, we plan to apply the method to other semantic categories using data from a variety of social media platforms. The analysis in [21] was based on the posts written by a large sample of English-speaking Facebook users over a period of more than three years between 2009 and 2012, and we restricted our analysis to the categories of positive and negative emotions defined by the LIWC. Although these two categories are negatively correlated, they are not opposite sides of the same scale. Heightened emotional arousal might cause users to express themselves with both categories at the same time.

A. Model Parameters

Table 2 and Fig. 1(a) show that rainfall is a valid instrument for both categories of positive and negative emotion (reprinted from [21, Fig. 2A]). That is, it predicts enough of the variability of the content posted that it allows us to obtain reliable estimates of influence with our method. Table 3 and Fig. 1(b) show statistically significant estimates γ of contagion (reprinted from [21, Fig. 2B]). In particular, a person's post in one semantic category can cause friends to generate one to two additional posts in the same category (see Section III-F). Also, an increase in the usage of positive (resp., negative) emotion words by an individual inhibits the usage negative (resp., positive) emotion words by her social contacts.

B. Additional Tests

Since we would expect that friends' future expression does not predict a person's current semantic expression, we can consider the following placebo model:

$$\bar{y}_g^{(c)}(t) = \theta(t) + \bar{f}_g + \beta\bar{x}_g(t) + \gamma\bar{Y}_g^{(c')}(t+\delta) + \bar{\epsilon}_g(t) \qquad (11)$$

where friends' future usage of category c' appears as an explanatory variable. We need to choose a lag of δ days in order to break the correlation between friends' present rainfall $\bar{X}_g(t)$ and future rainfall $\bar{X}_g(t+\delta)$. We can then estimate the model via 2SLS regression using friends' future rainfall $\bar{X}_g(t+\delta)$ as the instrument, and we would expect not to find statistically significant estimates of γ. In [21], we set $\delta = 30$ days and we found statistically insignificant estimates of γ for all considered models.

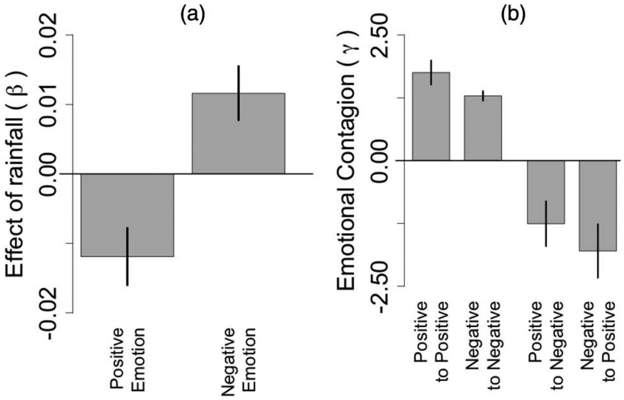
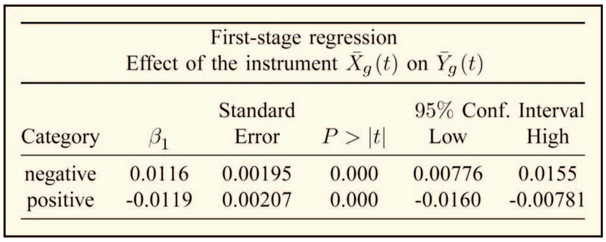


Fig. 1. ***(a) Effect β_1 of the instrument (friends' rainfall) on the endogenous explanatory variable (friends' positive and negative expression), from the first-stage regression. (b) Estimate of emotional contagion γ, from the second-stage regression. Vertical bars represent 95% confidence intervals. Reprinted from [21, Fig. 2].***

Table 2 Estimates of the Coefficient β_1 (With Additional Statistics and 95% CI) for the First-Stage Regression of (7) for the Categories of Positive and Negative Emotion. *p* Values Smaller Than 0.05 Reject the Null Hypothesis of Zero Coefficient. The Kleibergen-Paap *rk* LM Statistics Reject the Null Hypothesis That the Regression Is Underidentified [57]. The Cragg-Donald Wald *F* Statistics Exceed the Critical Thresholds Suggested by Stock and Yogo [58] to Ensure the Instruments Are not Weak. All Statistics Are Robust to Heteroskedasticity, Autocorrelation, and Clustering. Reprinted From Tables 6 and 7 of the Supplemental Appendix to [21].



First-stage regression
Effect of the instrument $\bar{X}_g(t)$ on $\bar{Y}_g(t)$

		Standard		95% Conf. Interval	
Category	β_1	Error	$P > \lvert t \rvert$	Low	High
negative	0.0116	0.00195	0.000	0.00776	0.0155
positive	-0.0119	0.00207	0.000	-0.0160	-0.00781

To test whether our estimates of influence are driven by people writing posts about the weather (a situation that would change our interpretation of the results), in [21], we considered a meteorological glossary supplied by the National Oceanic and Atmospheric Administration (NOAA),[3] and for each i and t, we defined $w_i(t)$ as the fraction of posts of subject i on day t containing a meteorological word. We consider the following version of model (1):

$$y_i^{(c)}(t) = \theta(t) + f_i + \beta_{c',c}x_i(t) + \delta w_i(t) + \gamma_{c',c}\frac{1}{\delta_i(t)}\sum_j a_{i,j}(t)y_j^{(c')}(t) + \epsilon_i(t)$$

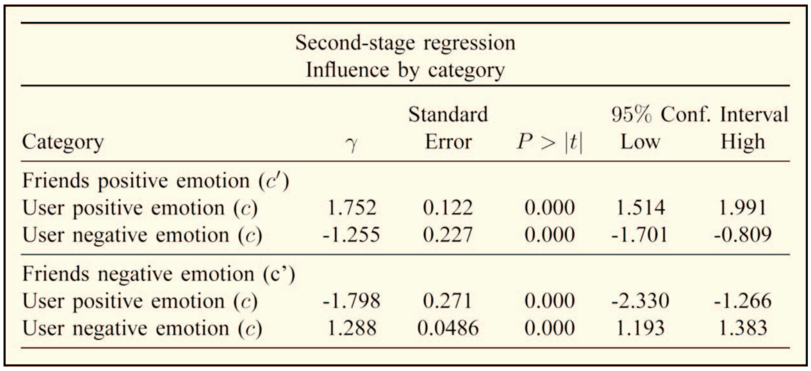
and its aggregated version

$$\bar{y}_g^{(c)}(t) = \theta(t) + \bar{f}_g + \delta\bar{w}_g(t) + \beta\bar{x}_g(t) + \gamma\bar{Y}_g^{(c')}(t) + \bar{\epsilon}_g(t) \qquad (12)$$

where $\bar{w}_g(t)$ is the average of $w_i(t)$ over all people in city g. The model is estimated via 2SLS regression, using $\bar{X}_g(t)$ as the instrument. Our results showed that when we control for weather-related words, the estimates of the influence coefficient γ for model (12) were unchanged with respect

[3] http://www.erh.noaa.gov/box/glossary.htm

Table 3 Estimates of the Coefficient γ (With Additional Statistics and 95% CI) for the Second-Stage Regression of Equation (8), When $c = c'$ and $c \neq c'$ for the Semantic Categories of Positive and Negative Emotion. p Values Smaller Than 0.05 Reject the Null Hypothesis of Zero Coefficient. Reprinted From Tables 5, 6, 7, and 8 of the Supplemental Appendix to [21]



Second-stage regression
Influence by category

Category	γ	Standard Error	$P > \lvert t \rvert$	95% Conf. Interval Low	95% Conf. Interval High
Friends positive emotion (c')					
User positive emotion (c)	1.752	0.122	0.000	1.514	1.991
User negative emotion (c)	-1.255	0.227	0.000	-1.701	-0.809
Friends negative emotion (c')					
User positive emotion (c)	-1.798	0.271	0.000	-2.330	-1.266
User negative emotion (c)	1.288	0.0486	0.000	1.193	1.383

to those for the original model (2). This suggests that the influence estimate is not driven by people writing posts about the weather.

V. DISCUSSION

In [21], we proposed a rigorous method based on mathematical modeling and instrumental variable regression to detect and quantify contagion of semantic expression in online social networks using observational data. First, our method allows us to determine what semantic categories are susceptible to peer influence between social contacts. In particular, we showed that a person's post expressing positive or negative emotion can cause his or her friends to generate one to two additional posts expressing the same emotion. Second, it allows us to estimate a signed relationship between different categories, characterizing how an increase in the usage of a semantic category by an individual alters the usage of another by her social contacts. Third, our model allows us to quantify the cumulative effect that a person has on all her social contacts.

One potential concern is the instrument's weakness [42]; rainfall has only a small effect in our analysis, but this does not harm the validity of our conclusions because it is the precision, and not the size of the estimate, that matters. In the data set we used in [21], built from content posted by millions of users, even a small effect is statistically significant and robust to a multitude of statistical tests against instrument weakness.

Our method limits inference to influence between subpopulations (individuals in different cities). Drawing conclusions about influence within a subpopulation (individuals in the same city) using observational data requires either the identification of a valid instrument or the definition of a different approach. This is an avenue of future research.

There are, of course, some limitations in inferring causality from observational data, and robust instruments may not always be available. Our model provides an alternative method when a large scale experiment is infeasible and researchers must rely on observational data. In an experiment, one would directly control the state of some people in order to track changes in their friends' outcomes (semantic expression, in our case). With the proposed approach, which constitutes a "natural experiment," the instrument (rainfall, in our case) constitutes a source of variation that affects some people directly (those experiencing it) but can predict changes in their social contacts who do not directly experience it. Moreover, our method can be easily applied to massive data sets (thanks to aggregation), and allows us to perform multiple analyses regarding several outcomes.

We advocate for the involvement of the engineering community in the development of nonexperimental methods of causal inference. On the one hand, it is an open question how methods based on instrumental variable regression generalize to different contexts (especially contagion within a population) and how to build instruments in a systematic way. On the other hand, although instrumental variables might provide interesting answers, researchers should also develop and propose alternative techniques. ■

REFERENCES

[1] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2009, pp. 497–506.

[2] S. Goel, D. J. Watts, and D. G. Goldstein, "The structure of online diffusion networks," in *Proc. 13th ACM Conf. Electron. Commerce*, 2012, pp. 623–638.

[3] N. A. Christakis and J. H. Fowler, "The spread of obesity in a large social network over 32 years," *New England J. Med.*, vol. 357, no. 4, pp. 370–379, 2007.

[4] N. A. Christakis and J. H. Fowler, "The collective dynamics of smoking in a large social network," *New England J. Med.*, vol. 358, no. 2, pp. 2249–2258, 2008.

[5] J. H. Fowler and N. A. Christakis, "Dynamic spread of happiness in a large social network: Longitudinal analysis of the Framingham Heart Study social network," *British Med. J.*, vol. 338, pp. 23–27, 2009.

[6] J. N. Rosenquist, J. H. Fowler, and N. A. Christakis, "Social network determinants of depression," *Molec. Psychiatry*, vol. 16, no. 3, pp. 273–281, 2011.

[7] S. Aral and D. Walker, "Identifying influential and susceptible members of social networks," *Science*, vol. 337, no. 6092, pp. 337–341, 2012.

[8] L. Muchnik, S. Aral, and S. J. Taylor, "Social influence bias: A randomized experiment," *Science*, vol. 341, no. 6146, pp. 647–651, 2013.

[9] S. Goel and D. G. Goldstein, "Predicting individual behavior with social networks," *Marketing Sci.*, vol. 33, no. 1, pp. 82–93, 2013.

[10] C. F. Manski, "Identification of endogenous social effects: The reflection problem," *Rev. Econ. Studies*, vol. 60, no. 3, pp. 531–542, 1993.

[11] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annu. Rev. Sociol.*, vol. 27, pp. 415–444, 2001.

[12] J. H. Fowler and N. A. Christakis, "Cooperative behavior cascades in human social networks," *Proc. Nat. Acad. Sci.*, vol. 107, no. 12, pp. 5334–5338, 2010.

[13] D. Centola, "The spread of behavior in an online social network experiment," *Science*, vol. 329, pp. 1194–1197, 2010.

[14] D. Centola, "An experimental study of homophily in the adoption of health behavior," *Science*, vol. 334, pp. 1269–1272, 2011.

[15] J. Guillory *et al.*, "Upset now? Emotion contagion in distributed groups," in *Proc. ACM SIGCHI Conf. Human Factors Comput. Syst.*, 2011, pp. 745–748.

[16] R. M. Bond *et al.*, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, no. 7415, pp. 295–298, 2012.

[17] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proc. Nat. Acad. Sci.*, vol. 111, no. 24, pp. 8788–8790.

[18] V. Goel, "Facebook tinkers with users emotions in news feed experiment, stirring outcry," *New York Times*, Jun. 29, 2014. [Online]. Available: http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html

[19] M. Wood, "OkCupid plays with love in user experiments," *New York Times*, Jul. 28, 2014. [Online]. Available: http://www.nytimes.com/2014/07/29/technology/okcupid-publishes-findings-of-user-experiments.html

[20] M. Schroepfer, "Research at Facebook," Oct. 2, 2014. [Online]. Available: http://newsroom.fb.com/news/2014/10/research-at-facebook

[21] L. Coviello *et al.*, "Detecting emotional contagion in massive social networks," *PLoS ONE*, vol. 9, no. 3, DOI: 10.1371/journal.pone.0090315.

[22] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC." [Online]. Available: http://www.liwc.net

[23] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, *The Development and Psychological Properties of LIWC 2007*. Austin, TX, USA: LIWC, 2007.

[24] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Social Psychol.*, vol. 29, no. 1, pp. 24–54, 2010.

[25] M. R. Mehl and J. W. Pennebaker, "The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations," *J. Pers. Social Psychol.*, vol. 84, no. 4, pp. 857–870, 2003.

[26] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annu. Rev. Psychol.*, vol. 54, no. 1, pp. 547–577, 2003.

[27] J. Graham, J. Haidt, and B. A. Nosek, "Liberals and conservatives rely on different sets of moral foundations," *J. Pers. Social Psychol.*, vol. 96, no. 5, pp. 1029–1046, 2009.

[28] J. D. Angrist, G. W. Imbens, and D. B. Rubin, "Identification of causal effects using instrumental variables," *J. Amer. Stat. Assoc.*, vol. 91, no. 434, pp. 444–455, 1996.

[29] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on twitter," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 65–74.

[30] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2008, pp. 207–218.

[31] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in *Proc. 4th Int. AAAI Conf. Weblogs Social Media*, 2010, pp. 10–17.

[32] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter," in *Proc. 20th ACM Int. Conf. World Wide Web*, 2011, pp. 695–704.

[33] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proc. 21st ACM Int. Conf. World Wide Web*, 2012, pp. 519–528.

[34] S. Aral, "Commentary-identifying social influence: A comment on opinion leadership and social contagion in new product diffusion," *Marketing Sci.*, vol. 30, no. 2, pp. 217–223, 2011.

[35] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proc. Nat. Acad. Sci.*, vol. 106, no. 51, pp. 21544–21549, 2009.

[36] J. D. Angrist and A. B. Krueger, "Empirical strategies in labor economics," *Handbook Labor Econ.*, vol. 3, pp. 1277–1366, 1999.

[37] D. Card, "The causal effect of education on earnings," *Handbook Labor Econ.*, vol. 3, pp. 1801–1863, 1999.

[38] G. M. Imbens and J. M. Wooldridge, "Recent developments in the econometrics of program evaluation," Nat. Bureau Econ. Res., no. w14251, 2008.

[39] J. R. Kling, J. B. Liebman, and L. F. Katz, "Experimental analysis of neighborhood effects," *Econometrica*, vol. 75, no. 1, pp. 83–119, 2007.

[40] J. Morduch, "The microfinance promise," *J. Econ. Literature*, vol. 37, no. 4, pp. 1569–1614, 1999.

[41] J. Bound, D. A. Jaeger, and R. M. Baker, "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak," *J. Amer. Stat. Assoc.*, vol. 90, no. 430, pp. 443–450, 1995.

[42] D. O. Staiger and J. H. Stock, "Instrumental variables regression with weak instruments," Nat. Bureau Econ. Res., no. 151, 1994.

[43] M. Stubbs, *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford, U.K.: Blackwell, 1996.

[44] M. Coulthard, Ed., *Advances in Written Text Analysis*. London, U.K.: Routledge, 2002.

[45] N. Fairclough, *Analysing Discourse: Textual Analysis for Social Research*. London, U.K.: Routledge, 2003.

[46] T. Joachims, *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Norwell, MA, USA: Kluwer, 2002.

[47] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text classification using string kernels," *J. Mach. Learn. Res.*, vol. 2, pp. 419–444, 2002.

[48] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2010, pp. 841–842.

[49] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proc. 17th Int. ACM Conf. World Wide Web*, 2008, pp. 91–100.

[50] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguist., Human Lang. Technol.*, 2011, pp. 151–160.

[51] R. Wardhaugh, *An Introduction to Sociolinguistics*. New York, NY, USA: Wiley-Blackwell, 2006.

[52] S. Dubois and B. Horvath, "Let's tink about dat: Interdental Fricatives in Cajun English," *Lang. Variat. Change*, vol. 10, no. 3, pp. 245–261, 1998.

[53] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *Proc. ACM SIGCHI Conf. Human Factors Comput. Syst.*, 2009, pp. 211–220.

[54] R. Xiang, L. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *Proc. 19th ACM Int. Conf. World Wide Web*, 2010, pp. 981–990.

[55] W. S. Robinson, "Ecological correlations and the behavior of individuals," *Amer. Sociol. Rev.*, vol. 15, no. 3, pp. 351–357, 1950.

[56] C. F. Baum, M. E. Schaffer, and S. Stillman, "Enhanced routines for instrumental variables/GMM estimation and testing," *Stata J.*, vol. 7, no. 4, pp. 465–506, 2007.

[57] F. Kleibergen and R. Paap, "Generalized reduced rank tests using the singular-value decomposition," *J. Econometrics*, vol. 127, pp. 97–126, 2006.

[58] J. H. Stock and M. Yogo, "Testing for weak instruments in linear IV regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, D. W. K. Andrews and J. H. Stock, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2005.

ABOUT THE AUTHORS

Lorenzo Coviello (Student Member, IEEE) received the Laurea degree (*magna cum laude*) in telecommunication engineering from the University of Padova, Padova, Italy, in 2008 and the M.S. degree in electrical and computer engineering from the University of California San Diego, San Diego, CA, USA, in 2013, where he is currently working toward the Ph.D. degree in electrical and computer engineering, under the advice of Prof. M. Franceschetti.

His research is focused on social networks, big data, and algorithms.

James H. Fowler received the B.A. degree (*cum laude*) from Harvard University, Cambridge, MA, USA, in 1992, the M.A. degree in international relations from Yale University, New Haven, CT, USA, in 1997, and the M.A. and Ph.D. degrees in government from Harvard University in 2001 and 2003, respectively.

He is a Professor of the Department of Political Science and School of Medicine, Medical Genetics Division, San Diego, CA, USA. His work lies at the intersection of the natural and social sciences, with a focus on social networks, behavior, evolution, politics, genetics, and big data.

Massimo Franceschetti (Senior Member, IEEE) received the Laurea degree (*magna cum laude*) in computer engineering from the University of Naples, Naples, Italy, in 1997 and the M.S. and Ph.D. degrees in electrical engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 1999, and 2003, respectively.

He is a Professor in the Department of Electrical and Computer Engineering, University of California San Diego, San Diego, CA, USA. He was a Postdoctoral Scholar at the University of California Berkeley, Berkeley, CA, USA. He has held visiting positions at the Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; the Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland; and the University of Trento, Trento, Italy. His research interests are in communication systems theory and include random networks, wave propagation in random media, wireless communication, and control over networks.

Dr. Franceschetti is an Associate Editor for Communication Networks of the IEEE Transaction on Information Theory (2009–2012) and has served as Guest Editor for two issues of the IEEE Journal on Selected Areas in Communication. He was awarded the C. H. Wilts Prize in 2003 for best doctoral thesis in electrical engineering at Caltech; the S.A. Schelkunoff Award in 2005 for best paper in the IEEE Transaction on Antennas and Propagation; a National Science Foundation (NSF) CAREER award in 2006; an Office of Naval Research (ONR) Young Investigator Award in 2007; and the IEEE Communications Society Best Tutorial Paper Award in 2010.