Computational detection of socioeconomic inequalities

Jacob Levy Abitbol

Numéro National de Thèse : 2020LYSEN001

THESE de DOCTORAT DE L'UNIVERSITE DE LYON

opérée par

l'Ecole Normale Supérieure de Lyon

Ecole Doctorale 512

Ecole Doctorale en Informatique et Mathématiques de Lyon

Spécialité de doctorat : Informatique

Discipline : **Informatique**

Soutenue publiquement le 09/01/2020, par :

Jacob LEVY ABITBOL

Computational detection of socioeconomic inequalities

Détection computationnelle des inégalités socioéconomiques

Devant le jury composé de :

Largeron, Christine	Professeur	Université J. Monnet, LHC	Rapporteuse
Sagot, Benoît	Chargé de Recherche	Université Paris 7, INRIA	Rapporteur
Besacier, Laurent	Professeur	Université J. Fourier, LIG	Examinateur
Carneiro Viana, Aline	Chargée de Recherche	INRIA Saclay	Examinatrice
Karsai, Márton	Professeur	CEU	Directeur de thèse
Fleury, Éric	Professeur	INRIA Paris	Co-encadrant de thèse

Computational Detection of Socioeconomic Inequalities

Jacobo LEVY ABITBOL

École Normale Supérieure de Lyon

A thesis submitted to the Laboratoire Informatique du Parallélisme in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

October 2019

Abstract

Socioeconomic inequalities drag down economic growth and hamper social cohesion. Their detection is therefore a necessary key step in the search of solutions addressing them.

In this dissertation, we address this challenge by exploiting the increasingly large collection of societal datasets to fuel novel computational pipelines able to generate fine-grained and accurate socioeconomic mappings. The updated and upscaled picture of society that is generated in doing so can be later used as input for the development of new critically needed socioeconomic policies. The technical focus of this dissertation consists then on both the constructions of datasets rich enough to expose these underlying correlations and the design of machine learning models able to generate accurate estimations on socioeconomic status both at the individual and neighborhood level.

To do so, we first collected and built several independent datasets enabling the fine grained description of France in terms of socioeconomic status, sociolinguistic information, social network and aerial imagery. The mining of the collected information enabled us first to analyze the underlying entanglements linking these variables to one another and later to use these interactions to build reliable predictors. These systems are built on techniques drawn from the deep learning and representation learning communities and allow for the reliable prediction of socioeconomic status to later inform developmental policies and gain insights on the way inequalities may rise in societies.

Résumé

Nous vivons une période marquante: pour la première fois, nous sommes conscients des enjeux de notre temps, nous produisons suffisamment de données pour en fournir une description complète et nous disposons d'algorithmes raisonnablement optimaux pour les traiter. Au centre de ce carrefour, une nouvelle discipline, la science sociale computationelle, profondèment imprégnée des avances en intelligence artificielle et en algorithmique, vient se dresser comme une sphère de connaissance à part entière.

Cette thèse s'inscrit dans cet élan et cherche à fournir des éléments de compréhension à la problématique des inégalités socioéconomiques en traitant des données massives, notamment issues de réseaux sociaux en ligne et de l'observation de l'environnement urbain. Ainsi, les contributions principales de cette série de travaux sont centrées autour de 1) l'étude des dépendances spatiales, temporelles, linguistique et du réseau liées aux inégalités et 2) l'inférence du statut socioéconomique à partir de ces signaux multimodaux.

Le contexte dans lequel cette série de travaux est inscrite est double. D'un côté, nous cherchons à fournir aux chercheurs et aux éléments du pouvoir décisionnel des outils qui leur permettront d'obtenir une image plus fine et détaillée de la répartition de richesse dans le pays dans le but qu'ils puissent adopter des stratégies portant à la résolution de deux défis de notre temps: la pauvreté et les inégalités socioéconomiques. De l'autre nous cherchons nous même à fournir des éléments de réponse aux questions posées par les sciences sociales qui se sont avérées trop intractable pour être abordées sans le volume et la qualité de données nécessaires.

Acknowledgements

This work could not have been achieved without the continued support of my parents and sister as well as my friends and lab peers Sam, Sebastian, Marija, Rodrigo, Misha, Esteban and Jazmin, any words I could offer would only fail to grasp the full extent of your help. I am also grateful to my supervisors Márton Karsai and Éric Fleury, who offered undubbious backing, regular encouragement and thorough evaluations of my ongoing work.

I would like to thank Benoît Sagot and Christine Largeron for accepting to be examiners of my thesis, as well as Aline Carneiro Viana and Laurent Besacier for taking part as jury members.

I additionally would like to thank Jean-Philippe Magué, Jean-Pierre Chevrot and Djamé Seddah for their precious contributions to our work on Twitter, as well as Simon Delamare and Dominque Ponsard for their unwavering IT and administrative help over the years. To all of you, thank you.

Contents

List of Figures

List of Tables

Introduction en français

Nous vivons une période particulièrement marquante de notre Histoire: pour la première fois, nous sommes conscients des enjeux de notre temps, nous produisons suffisamment de données pour en fournir une description complète et nous disposons d'algorithmes raisonnablement optimaux pour les traiter. Au centre de ce carrefour historique, une nouvelle discipline, la science sociale computationelle, profondément imprégnée des avances en intelligence artificielle et en algorithmique, vient se dresser comme une sphère de connaissance à part entière. En tant que telle, elle vise l'appropriation des questions issues de sciences sociales, en particulier celles qui se sont avérées intractables pour être résolues par leurs méthodes, pour en fournir des réponses axées sur le traitement de données massives.

Cette thèse s'inscrit dans cet élan et cherche à fournir des éléments de compréhension à la problématique des inégalités socioéconomiques en traitant des données massives, notamment issues de réseaux sociaux en ligne et de l'observation de l'environnement urbain.

Premièrement, nous débutons notre étude en analysant le lien entre le style linguistique des individus, leurs réseaux sociaux, leur activité spatio-temporelle et leur statut socioéconomique. Les données utilisées lors de ces étude sont issues du plus grand corpus de tweets (*Twitter*) en France collecté jusqu'à présent. Afin de structurer ce dernier, on introduit une série d'algorithmes permettant 1) la localisation du domicile des utilisateurs à partir de leurs traces de géolocalisation 2) l'obtention de leur statut socioéconomique en plaçant chaque individu au sein d'une grille socioéconomique de haute-résoluttion à partir de leur localisation et 3) la reconstruction du réseau social des utilisateurs via la définition d'une intéraction sociale à partir de l'ocurrence d'une mention mutuelle (bidirectionelle) entre un couple d'usagers. Cette approche permet ensuite de détecter des corrélations socioéconomiques à partir du recouvrement engendré par ces trois couches de données.

Nous poursuivons nos études ensuite en construisant des algorithmes sur la base des corrélations entre classe sociale, language et réseau détectées au préalable afin de prédire le statut socioéconomique des utilisateurs Twitter contenus dans notre jeu de

données. Au-delà de la construction d'une méthodologie solide permettant ce genre d'inférence, nous abordons de-même la question du biais engendré par la dépendence vis-à-vis d'une source de données particulières. Plus concrètement, plutôt que de définir le statut socioéconomique des utilisateurs uniquement à partir de l'adresse de leur logement, nous élargissons le contexte analysé en assignant la classe sociale à partir de la catégorie socio-professionelle et de l'observation de l'environnement urbain du logement des usagers (Chapitre 3).

En parallèle, dans le chapitre 4, on focalise notre attention sur les corrélations masquées qui existent entre la classe sociale de chaque individu et son réseau social. Pour cela, on développe un algorithme d'inférence du premier à partir du dernier via la mise en place d'un modèle d'aprentissage basé sur des réseaux convolutionnels appliqués aux graphes (*Graph Convolutional Network, GCN*). Ce dernier est ensuite utilisé pour appréhender le recouvrement entre attributs individuels de chaque noeud du réseau et le réseau lui-même.

Finalement, dans le chapitre 5, on clôt notre étude en s'interrogeant sur la question de la prédiction du statut socioéconomique d'une fine aire urbaine en croisant images satellites et intelligence artificielle pour extraire de l'information économique automatiquement à partir de l'image aérienne de la zone en question. Pour cela nous collectons des données aériennes de toutes les aires urbaines françaises et développons un algorithme d'aprentissage machine capable de réaliser ces prédictions pour cinq villes françaises avec des précisions comparables à l'état de l'art. Nous mettons ensuite en valeur les éléments décisionnels employés par notre réseau de neurones, en analysant et groupant les activations de ce dernier en fonction du tissu urbain sous-jacent.

Le contexte dans lequel cette série de travaux est inscrite est double. D'un côté, nous cherchons à fournir aux chercheurs et aux éléments du pouvoir décisionnel des outils qui leur permettront d'obtenir une image plus fine et détaillée de la répartition de richesse dans le pays dans le but qu'ils puissent adopter des stratégies portant à la résolution de deux défis de notre temps: la pauvreté et les inégalités socioéconomiques. De l'autre nous cherchons nous même à fournir des éléments de réponse aux questions posées par les sciences sociales qui se sont avérées trop intractable pour être abordées sans le volume et la qualité de données nécessaires.

Chapter 1

General Introduction

Uncovering the patterns characterizing human behavior has become one of the main challenges of modern-day social sciences [162]. This claim becomes all too evident when one starts considering the numerous phenomena that are driven at their core by the way individuals behave and interact with one another. The successful comprehension of the spread of epidemics [178], the roots and consequences of socio-political polarization [42], the design of transportation systems [12] as well as the weaving of our own social networks [193] are all reliant upon the detailed understanding of how we behave individually and collectively.

Despite the inherent edge that is gained by studying human dynamics, previous research in this direction has been plagued by the difficulty of gathering observational data at scale and over time [132], inherently begetting a scarcity of reliable mechanistic models on human behaviour and society. This initial setback has however been overcome in recent years due to three major feats: the successful development and fast widespread adoption of new technologies, their subsequent ease in mining human behaviour and the surge in increasingly optimal ways of processing and analyzing the generated data [155].

Just as modern day deterrence policies are based upon the so-called nuclear triad, this data triad is underpinning a paradigm shift in the social sciences, that keep moving away from incomplete and a posteriori descriptions of individual and collective behaviour and tilting towards a more complete data-fueled modeling of observed human behaviour. At the core of this shift lies the burgeoning field of computational social science [132, 64] which aims to leverage the sea of information generated by these new technologies to address longstanding questions in the social sciences.

While one might be tempted to underscore the importance of each of the individual factors behind this data triad, their effects are deeply intertwined and cannot

be thought of in isolation from one another: the Internet, mobile phones and online social networks have radically reshaped the means by which we communicate and interact with one another at a global scale [155]. The social interactions these technologies enable leave in turn, by design, massive digital traces of individual and collective habits. Mobility, consumption patterns, political leanings, personality traits are among all the digital crumbs left by the widespread use of these technologies. Their inference from raw data is nevertheless far from trivial: users of these services are not required to provide self-defining attributes and the recorded data streams are rarely well-structured or parsed at all. This caveat has typically hindered the use of off-the-shelf algorithms to derive information of the sort and would probably have limited the insights gained from these large volumes data where it not for the advent and refinement of machine and deep learning models.

Traditionally, the use of machine learning (ML) models necessitated careful engineering and considerable domain expertise to design a feature extractor able to turn raw data such as pixel values in an image or words in a sentence into a suitable feature vector from which the classifier could detect or classify patterns in the input. Once introduced, deep learning (DL) radically challenged this approach by learning the representations of the data best suited for classification directly from the data itself [133]. By composing several non-linear modules (layers), DL models are able to transform the input sequentially into a more complex and abstract representation. Effectively, their architecture is generally composed of multiple layers of simple modules stacked one after the other. Most of these modules feature weights, which optimal values need to be learned and many of them compute non-linear input-output mappings. Each stacked module can be thought of as a series of filters augmenting the features of the input that are most important for discrimination and discard useless variations. When multiple non-linear layers are stacked (hence deep learning), a neural network can implement very complex functions of its inputs. If trained on a set of images for instance, it becomes able to sense minute details in each image while at the same time discarding large irrelevant differences between samples such as the background, pose, lighting and surrounding objects. The heavy parametrization that enables DL models to learn these intricate functions relies however upon 1) the right setup of the optimization routine to determine its parameters (weights) as well as 2) the availability of a large enough collection of samples from which the model can learn representations generalizable to unseen data. Furthermore, the same volume of data that enables the models' finetuning as well as the effective learning of these representations is also key in improving the design of the models themselves as new architectures are developed to make a more efficient use of the signals contained within the data.

Data is of course not the single actor behind this success story, which has actu-

ally hinged upon the introduction of several key design and implementation features: proof that multilayer architectures could be trained by stochastic gradient descent (SGD) provided that the non-linear models were smooth functions of their inputs and their internal weights [192], pre-training of bottom layers [89, 15], introduction of ReLU (Rectified Linear Units) non linearities speeding up learning in networks with many layers (deep networks) [68], the advent of fast graphics processing units (GPUs) [186] and design of regularization techniques such as dropout [208] and data augmentation routines [179] can all be credited in part for the current importance of deep learning.

Nevertheless the reasons why these methods are particularly well poised to discovering intricate structures in high-dimensional data, it is hardly arguable that they have dramatically improved the state-of-the-art in tasks ranging from speech recognition [159, 88] to visual object recognition [206] and object detection [65], with DL based solutions being deployed in every major software company [133] and new models being developed to deal with machine learning related tasks on other structures (graphs [176, 76, 116, 80], sequences [90, 212, 39]) and domains (biomedicine [139, 223, 107], chemistry [51, 180]).

Social sciences have accordingly not been exempted from benefiting from deep learning based advances. Neural language models [14, 211, 69] have been extensively used to analyze social and linguistic based phenomena such as, for instance, quantifying semantic change and revealing statistical laws of semantic evolution [78], detecting the ideological bias of news articles [124] or analyzing the spread of true and false news online [226] among others. Similarly, convolutional neural networks have been used in urban settings to explore the dynamics of urban change by determining which characteristics predict neighborhood improvement [164] or by examining how the perception of safety affects the liveliness of neighborhoods [45]. Such works are paving the way for the use of social representations learned from raw data as means of answering hitherto intractable social sciences questions.

In light of this fact, the core argument of this dissertation is that one particular social phenomenon, socioeconomic inequalities, can be gainfully studied by means of the above. Therein lies the main motivation of this thesis that sets out to collect and combine large datasets enabling **1) the study of the spatial, temporal, linguistic and network dependencies of socioeconomic inequalities** and **2) the inference of socioeconomic status (SES) from these multimodal signals**. This task is one worthy of study as previous research endeavours have come short of providing a complete picture on how these multiple factors are intertwined with individual socioeconomic status and how the former can fuel better inference methodologies for the latter. As I will endeavour to demonstrate in what follows, the

study of these questions is important, as much is still unclear about the root causes of SES inequalities and the deployment of ML/DL solutions to pinpoint them is still very much in its infancy.

Understanding how this phenomenon came to be and whether it should be solved at all is of utmost importance to fully appreciate the contributions that we'll here be presented. What follows will introduce the reader to these notions and clarify the starting point of this research within the deep learning and machine learning for social good communities.

A brief introduction to computational socioeconomics

A coarse historical overview of socioeconomic inequalities

Ever since the development of agriculture and herding, socioeconomic inequality has been a defining feature of our civilization. Prior to that, humans have been shown to live in small bands, numbering several dozens individuals, that mainly foraged and hunted as a means of subsistence [81]. These societies are thought to actually have been determined by low-levels of resource inequality and a strong egalitarian, sometimes aggressive ethos, where inequality was kept at bay by due to logistical and infrastructural constrains. Indeed, the prevalence of a nomadic lifestyle with no pack animals limited the accumulation of material possessions while the small size and flexible composition of human groups left little room for establishing stable asymmetric relationships beyond those related to age and gender [205].

Around 12,000 years ago however, in multiple locations, humans started cultivating crops and developed agriculture. Contrary to hunting and foraging, this new technique was a reliable way to yield a predictable food supply, allowing for the production of food surplus. Food surplus led to specialization of labor as well as population growth and concentration enabling, as farming got more efficient, for cities and later states to emerge and what we call civilization to begin. It also generated a transition into new forms of social organization that eroded forager egalitarianism and replaced it with durable hierarchies and disparities in income and wealth [184].

The formation of states did little to flatten these emerging disparities, as these new forms of organization established structures of power and force, skewing access to income and wealth even further. Political inequality reinforced and amplified economic inequality and, for most of the agrarian period, the state enriched a few at the expense of the many. Even when more benign institutions took over and promoted more vigorous economic development, they continued to sustain high inequality. Urbaniza-

tion, commercialization, innovation in the technological and financial sector, global trade and finally industrialization gradually generated high returns for the holders of capital. Even as economic structures, social norms and political systems changed, income and wealth inequality remained high or found new ways to grow [196].

In developed countries, only recently the chaos of the first and second world wars and the Depression have disrupted this pattern. High taxes, inflation, bankruptcies and the growth of sprawling welfare states caused wealth to shrink dramatically, and ushered in a period in which both income and wealth were distributed in relatively egalitarian fashion. But the shocks of the early 20th century have faded and wealth is now reasserting itself [177]. Indeed, income disparities have been growing in all countries for which data is available since the eighties: if in 2010, 388 persons owned as much private net wealth as the poorer half of humanity, 62 billionaires were required to do so in 2015 and 26 in 2019, carrying the burden of socioeconomic inequality to the present day [82].

Am I My Brother's Keeper?

While abundant data has been compiled to confirm the rise of income inequality in modern developed countries (see for instance Figure 1-1), the question of whether income and wealth inequality as such have a negative effect on the lives of individuals is still an open one [196]. There is nevertheless some evidence to suggest so. Leaving aside ethical objections to the skewed distribution of resources (which are undoubtedly worthy of consideration), approaches to tackle these questions have varied across social sciences. On the one hand, economists have focused on the relation between socioeconomic inequality and economic growth and shown that higher levels of inequality are actually linked to lower rates of growth while lower disposable income inequality has been found to lead to faster growth and longer growth periods. Recent research has even suggested that while moderate wealth inequality may enables cooperation within a group, extreme inequality actually ends up preventing [84]. On the other hand, sociologists and political scientists have linked higher inequality in developed countries to less inter-generational economic mobility [43] and increasing political polarization [173] while in developing countries, certain kinds of income inequality have been shown to heighten the possibility of internal conflicts and civil wars [22]. These series of ideas were probably best summarized by U.S Supreme Court Justice Louis Brandeis: "We can either have democracy in this country or we can have great wealth concentrated in the hands of a few, but we can't have both."

Despite this accruing body of evidence, other studies have actually downplayed inequalities overall importance in democratic societies in the developed world. For instance, Scheidel [196] pointed out to the occurrence of economic equality without strong democratic governments while Scheve and Stasavage showed that the democ-

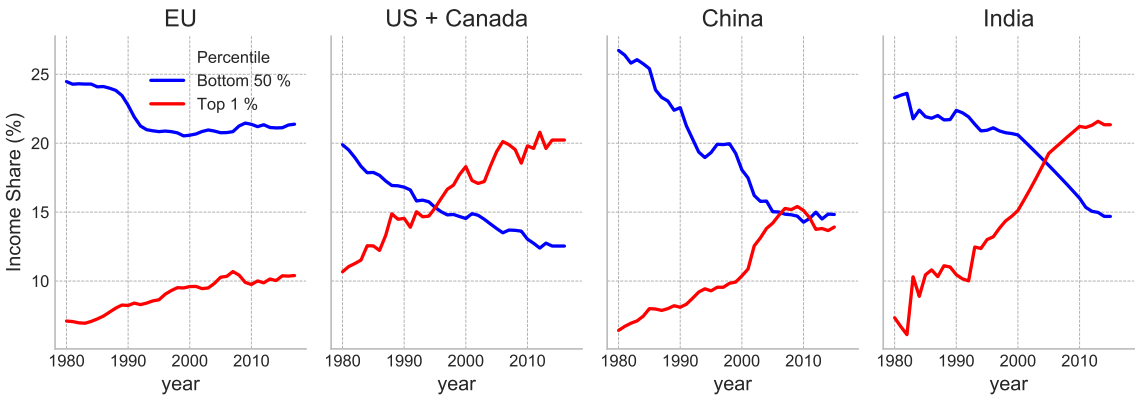


Figure 1-1: Share of national income held by the bottom 50 % (blue) and top 1 % (red) of earners. The rise of income inequality varies greatly across world regions being the lowest (albeit occurring) in Europe and among the highest in India. Data retrieved from the World Wealth and Income Database [127].

ratization of the West didn't constrain material inequality [197].

In spite of the significant amount of effort devoted to the study of inequality, never mind its effects as a whole in societies, the actual issue that is addressed when trying to uncover the mechanisms and consequences of inequalities is without a doubt the prevalence of poverty, said otherwise, as H. Frankfurt stated it, "[the issue] is not that the incomes [...] are widely unequal. It is, rather, the fact that too many people of our people are poor" [58]. To be sure, both notions are highly entangled, and the nature of that entanglement actually varies with the guiding principles chosen to define poverty.

Indeed, it might actually look surprising to learn that despite this growth in absolute inequality, absolute poverty has over the same period declined dramatically, with 1.5 billion people having overcome it in the span of 50 years at the start of the millennium [187]. Furthermore, this astounding achievement pales in comparison to the current goal set by the World Bank to eradicate poverty by 2030 considered by some to be fully within reach. The choice of metric gives a more complete view on how both these trends can actually co-occur, as absolute poverty is measured as the number of people living on less than 1.90 $ per day [9]. This in turn makes the notion of absolute poverty difficult to export to developed countries, which explains why in these cases, poverty rates are determined either by a different disposable income threshold (25 100 $ for the US), or through a relative poverty measure as most European countries do, by considering people to be poor when they have less income and opportunities than other individuals living in the same society (for example if they earn less than 60% of the median income [171]), inducing relative poverty to become actually a metric of inequality. Therefore, when accounting for these new variables,

the trend seems to be reversed and hint towards to a rise in relative poverty [187] mimicking the surge in absolute inequality discussed prior.

Because of their effects on society and modern economies as a whole, these two intertwined issues have recently become two of the seventeen goals for sustainable development set by the United Nations [167] and therefore must attract policies and legislation that help curb their effects and eventually render them a curse of the past. To ensure their success though, decision-makers must rely on two commodities to inform their actions: reliable societal data and robust statistical models able to extract meaningful information on the processes generating these phenomena.

The Solution to these Problems lies in Data

In order to understand the processes generating the above social phenomena, having reliable datasets on which to observe and test the underlying mechanisms thought to be at play is quintessential. Historically speaking, for most of human history, the best and sometimes only source to do so was the census. The first census dates back to 3800 BC in the city-state of Babylon where every six or seven years official censors counted the men, women, children, livestock, slaves, butter, milk, honey and vegetables in the kingdom [209]. The primary reason was to figure out how much food was needed to feed the population, but the figures also gave an idea of how many men were available for military service and how much they could be taxed without starving them. Ever since, as states consolidated and their bureaucracies expanded, their regularity and scope increased accordingly, nowadays covering information ranging from ethnic origins to language, occupation and income of each nations' citizens.

This historical continuity in our ability to gather meaningful mass-scale population information has however been severely disrupted in recent years due to two major trends: the abrupt and rapid adoption of new technologies, namely mobile devices and online services, as well as our increasing ability to process with ease the data generated by them [155].

These elements when taken together actually ensure that as the cost of storing data keeps decreasing and the ML/DL algorithms' performance keeps increasing, access to massive socioeconomic information and the ability to crunch it, though once only available to states or big companies, is increasingly becoming democratized. In doing so, not only is our ability to make better informed decisions regarding these social problems enhanced but we also become able to significantly improve our understanding of the causes that lie at their basis [132]. Before delving into the core of the works conducted during this thesis, we'll briefly see a shallow overview of some of the domain concepts we'll make use of.

Extremely Brief Machine Learning Overview

In the course of this dissertation, the reader will appreciate that most of the methods used or developed fit into one of three AI subfields, namely, computer vision, natural language processing or network embedding. To the untrained eye, each of these areas might seem to have developped in isolation from the others but as we'll here see that is far from the case.

In language modelling, one seeks to learn a representation mapping a discrete word into a continuous vector space. Traditionally, NLP systems would treat individual words as atoms, i.e, each word was represented by a single index in a long vocabulary list. For instance, in the widely used N-gram model [26], for a given sentence $w_{1:n}$ considered as a sequence of individual words $(w_i), i \in \{1, .., n\}$, the probability of a given word given the previous ones was estimated under the markovian assumption:

$$P(w_{i+1}|w_{1:i}) = P(w_{i+1}|w_{i-N:i})$$

by means of Maximum Likelihood Estimation. The obtained encodings are of course arbitrary, and fail to provide any meaningful information regarding the syntactical or semantical similarity that might exist between pair of words. A more insightful representation of words can be derived from the distributional hypothesis of Harris [83], which states that words in similar contexts have similar meanings. In practice, this involves the design of a vector space model the embeds every single word in a continuous vector space where words appearing in the same context, i.e co-appearing for example within a certain word distance from each other in a sentence, are mapped to nearby points (i.e embedded close to each other).

In practice, one straightforward way to leverage this principle is to learn word representations by means of a word-context matrix M where the value M_{ij} corresponds to some association measure between the word indexed by i and the context in which that word appeared indexed by j. To do this, we assume a corpus of words $w \in V_W$ and their contexts $v \in V_C$, where V_W and V_C are the word and context vocabularies. In this setting, words would come from an unannotated textual corpus of words $w_1, w_2, ..., w_n$ and the contexts for word w_i are the words surrounding it in an L-sized window $w_{i-L}, ..., w_{i-1}, w_{i+1}, ..., w_{i+L}$. $\#(w, c)$ is then the number of times the pair (w, c) appeared in the collection of observed words and context D while $\#(w) = \sum_{c' \in V_C} \#(w, c')$ and $\#(c) = \sum_{w' \in V_W} \#(w, c)$ are the number of times w and c occurred respectively in D. Every word in the vocabulary would then be modeled as a row either in M or some dimensionality-reduced matrix computed from M.

Neural language models [161, 140] pushed this concept further by proposing to rep-

resent every word as a dense vector computed by means of a neural network. Under this setting, each word $v \in V_W$ and each context $c \in V_C$ are respectively associated with a vector $\overrightarrow{v}, \overrightarrow{c} \in \mathbb{R}^d$ which entries are latent and thought of as parameters to be learned. To compute this dense representation, the skip-gram with negative-sampling (SGNS) training method (`word2vec`) in particular maximizes the dot-product between the vectors of frequently occurring word-context pairs, and minimizes it for random word-context pairs (negative examples hence the name negative sampling), i.e, samples for which the context is chosen at random for a given word resulting most likely in an unobserved pair (w, c). More formally, if we consider a word-context pair (w, c), the probability (w, c) came from the data D, $P(D = 1|w, c)$ is estimated as:

$$P(D = 1|w, c) = \sigma(\overrightarrow{w}.\overrightarrow{c}) = \frac{1}{1 + e^{-\overrightarrow{w}.\overrightarrow{c}}}$$

Conversely, the probably the pair (w, c) didn't occur in D is estimated as :

$$P(D = 0|w, c) = 1 - P(D = 1|w, c) = \sigma(-\overrightarrow{w}.\overrightarrow{c})$$

The SGNS objective function to maximize is then defined by:

$$L = \sum_{w \in V_W} \sum_{c \in V_C} \#(w, c)(\log\, \sigma(\overrightarrow{w}.\overrightarrow{c}) + k.\mathbb{E}_{c_N \sim P_D}\left[\log\, \sigma(-\overrightarrow{w}.\overrightarrow{c_N})\right])$$

where k is the number of negative samples and c_N is the sampled context drawn according the empirical unigram distribution: $P_D(c) = \frac{\#(c)}{|D|}$. The objective is hence trained using stochastic gradient descent over the observed pairs in the corpus D yielding, for every word, a low-dimensional dense vector that captures both syntactic and semantic word relationships. Notice that despite its impressive performance, SGNS actually relies on a shallow two layer neural network where the learned weights of the first and second layer factually correspond to look-up tables for each type of embedding $(\overrightarrow{w}, \overrightarrow{c})$. These learned latent representations are invaluable as they can later be used for instance for downstream classical supervised learning tasks such as predicting the political leaning of a given comment [183] or whether a given message expresses suicidal thoughts [75].

While learning these word representations is aided by the regularities governing language (grammatical rules, defined ordering of words in a sentence), extending this task to networks is far from a trivial task. A network can be generally thought of as a set of items, called nodes, with connections between them, called edges (or links or ties). Mathematically, it is represented by an undirected graph defined by a pair of sets $G = \{V, E\}$, where V is a set of N nodes/vertices $u_1, u_2, ..., u_N$ and E is a set of M edges linking pairs of elements from V. A node typically represents an entity, an

individual for example, while an edge represents a relation between two entities, such as a friendship between two individuals. This framework can be further extended to assign directionality (directed graphs) or weight to each edge (weighted graphs). In this setting, the learning of network representations implies the embedding of every node in the graph in a vector space whose geometry reflects the relationships between the nodes themselves, i.e neighbouring nodes should be projected close to each other in the embedding space.

A novel way to do so was introduced in a series of papers by Perozzi and collaborators [176] and Grover and collaborators [76] by re-purposing the aforementioned SGNS neural language model for social networks. At the core of this bridging lies the realization that both word frequency in natural language and node frequency in short random walks on scale free networks both follow a power-law distribution (Zipf's law [234] for the former). These series of models proceeded thus to explore the graph through a stream of short biased random walks which are thought of as short sentences and phrases in a special language. In doing so, each node is, as in language modeling, mapped to its latent representation that is in turn learned so as to maximize the likelihood of observing pairs of nodes co-occurring in random walks and minimize the likelihood of those not doing so. More formally, let $G = (V, E)$ be a given (unweighted, undirected) network and $f : V \rightarrow \mathbb{R}^d$ the function mapping from nodes to embedding we aim to learn. For every node in the network $u \in V$, we obtain its network neighborhood $N_S(u)$ through the random walk procedure previously introduced. We then seek to maximize the the log-probability L of observing a network neighborhood $N_S(u)$ for a node u conditioned on its feature representation, given by f:

$$\begin{aligned} L &= \sum_{u \in V} \log\, P(N_S(u)|f(u)) = \sum_{u \in V} \sum_{v \in N_S(u)} \log\, P(v|f(u)) \\ &= \sum_{u \in V} [-\log\, Z_u + \sum_{v \in N_S(u)} f(u).f(v)] \end{aligned}$$

as $P(v|f(u)) = \frac{\exp(f(u)f(v))}{\sum_{v' \in N_S(u)} f(u)f(v')}$. The Z_u normalization constant is furthermore computed via negative sampling, making the connection with the SkipGram architecture even clearer.

While these shallow architectures have proven themselves to be more than capable of generating meaningful representations for social networks and language, they have been generally found to lack sufficient representation power to be effectively used in computer vision tasks. For these, one particular type of deep neural network has been recursively shown to be much easier to train than deep networks with full connectivity between adjacent layers and to generalize much better than shallow ones. This is the

Convolutional Neural Network or ConvNet for short.

Just as regular deep networks, ConvNets are made of a sequential assembly of many layers. The structure of these layers is however not preserved as we delve deeper into the network: while deeper units can be fully connected layers, identical to the ones found in a standard multilayer neural network, its initial layers are comprised of one or more convolutional layers intertwined with subsampling or pooling layers [133]. This architecture is in fact designed to take advantage of the regularities in the 2D/3D structure of an input image by means of local connections and tied weights followed by some form of pooling which results in translation invariant features [134, 121]. By feats of this design, their resulting number of parameters is lessened with respect to equally deep fully connected networks, which in turn makes them faster to train and better at generalizing. These initial ConvNet architectures were subsequently improved by the introduction of network within network modules [213] permitting the computation of more abstract features for local patches of the input, residual connections [85] enabling the successful optimization of deeper networks by preventing the emergence of non-convexities in the loss landscape [144], dense blocks [99] alleviating vanishing gradient issues and strengthening feature propagation, and depthwise separable convolutions factorizing standard convolutions in previous architectures and turning them more efficient with little loss in accuracy [96].

Most recently, ConvNets successes in computer vision tasks have diffused to the network representation learning community: in a series of works [46, 51, 116, 80], ConvNets were generalized to work on arbitrarily structured graphs by relying upon graph convolutions known from spectral graph theory to define approximate smooth parametrized filters that are then used in a multi-layer neural network model. These models, which currently outperform previous SkipGram based network embedding models in standard benchmarks, will be covered in greater detailed in the main body of this thesis.

This brief summary is not meant to be taken as a detailed overview of any of three aforementioned areas but rather as a means for the reader to 1) understand how ML models devised for a particular task find their way into new disciplines and 2) get a sense of the general state-of-the-art techniques available at the outset of this work. Even if later chapters will go into greater detail regarding these methods, we nonetheless refer the reader to excellent surveys [24, 79, 188] for comprehensive overviews of these areas separately.

An improved understanding of society

Let us return to the matter at hand, socioeconomic inequalities. We have so far reviewed works that argue that not only is it not a new phenomenon but that it has also kept increasing steadily through periods of stability. Furthermore, in hopes of highlighting its importance, we have examined further research suggesting that absolute inequality is, as a whole, generally counterproductive for economic growth and deeply tied with social issues ranging from political polarization and low intergenerational economic mobility to relative poverty in developed countries (where most of the work we'll see is actually set). It may then be surprising to learn that in spite of all the attention this issue has attracted as well as all the drawbacks it has been linked to, economics has yet no accepted theory of inequality, i.e, no scientific understanding of the mechanisms and processes that engender it [40]. The mechanistic comprehension of this and other socioeconomic phenomena is critical as our inability to do so has so far hindered the forecasting and prevention of past economic crisis [13] as well as curbed the enactment of effective policies for economical development [147].

Rising data production, storage and processing capabilities are however tilting the scale. By fusing this updated and upscaled picture of society with the already stress-tested methods borrowed from statistical physics, learning theory and network science, researchers are effectively inducing a paradigm shift in the social sciences that greatly emphasizes the reliance upon quantitative description of social systems [62]. To be sure, there are numerous issues bedeviling this approach: a biased access to technology entails that the demographic coverage offered by these new sources is also biased. The degree to which complete answers can be provided to research questions is limited by our inability to fully control the experiments conclusion will be drawn from. Repeatability and reproducibility of observed patterns are curtailed both by the restrictions naturally set given the sensitive and sometimes private nature of the collected datasets as well as the bursty and heterogeneous that is known to characterize human behaviour [11].

Nevertheless, these limitations should not be thought of as barriers to using this quantitative approach but rather as research avenues of themselves, whose answers have, and will continue to ensure the robustness of observed trends, further shedding light upon the hitherto intractable problems at hand.

Thesis Outline

In light of the above, we set out to develop novel computational pipelines able to 1) detect population-scale socioeconomic correlations with respect to language, social

networks and urban environments and 2) make use of these correlations as building blocks for socioeconomic status inference algorithms. This dissertation discusses works conducted during this degree that address these challenges in the hope that they'll be used in further research or policy-planning to solve some of the societal issues of our time.

In Chapter 2, we show how key linguistic variables measured in individual Twitter streams are linked to user features such as socioeconomic status, location, time, and the social network of individuals. To do so we rely in a multi-layered dataset constructed from the largest Twitter sample collected within France and the finest-grained socioeconomic map of the country.

Chapter 3 builds upon these correlations to construct three different data collection and combination methods, each issued respectively from Twitter, LinkedIn and Google Maps, to first estimate and, in turn, infer the socioeconomic status of French Twitter users from their online semantics. We further establish links between the estimated social class of users and their mobility patterns.

Chapter 4 presents a theoretical framework aimed at further refining our understanding of correlations between one's network and socioeconomic status by jointly reconstructing an attributed network by means of a variational autoencoder relying on Graph Convolutional Networks. We further provide comparison with other state-of-the-art methods in this field.

Finally in chapter 5, we provide a separate Convolutional Neural Network based model able to predict socioeconomic status from aerial imagery at a 200m x 200m scale for five French cities. Given the importance of this task as well as its increasingly pervasive use by state agencies and NGOs, we further analyze the relation between the socioeconomic predictions yielded by the trained network and the underlying urban topology of each city.

Chapter 6 concludes with final thoughts, challenges and potential future work.

Chapter 2

Socioeconomic Correlations of Linguistic Patterns

This work was originally published in the Proceedings of The Web Conference [2] with myself as leading author

2.1 Introduction

The design of successful inference algorithms of individual socioeconomic status (SES) relies on the detailed understanding of the latent correlations linking SES to other facets of individual behaviour. To guide the former we delve into the study of the latter.

Communication is highly variable and this variability contributes to language change and fulfills social functions. Analyzing and modeling data from social media allows the high-resolution and long-term follow-up of large samples of speakers, whose social links and utterances are automatically collected. This empirical basis and long-standing collaboration between computer and social scientists could dramatically extend our understanding of the links between language variation, language change, and society.

Languages and communication systems of several animal species vary in time, geographical space, and along social dimensions. Varieties are shared by individuals frequenting the same space or belonging to the same group. The use of vocal variants is flexible. It changes with the context and the communication partner and functions as "social passwords" indicating which individual is a member of the local group [86]. Similar patterns can be found in human languages if one considers them as evolving and dynamical systems that are made of several social or regional varieties, overlapping or nested into each other. Their emergence and evolution result from their

internal dynamics, contact with each other, and link formation within the social organization, which itself is evolving, composite and multi-layered [120, 130].

The strong tendency of communication systems to vary, diversify and evolve seems to contradict their basic function: allowing mutual intelligibility within large communities over time. Language variation is not counter adaptive. Rather, subtle differences in the way others speak provide critical cues helping children and adults to organize the social world [113]. Linguistic variability contributes to the construction of social identity, definition of boundaries between social groups and the production of social norms and hierarchies.

Sociolinguistics has traditionally carried out research on the quantitative analysis of the so-called linguistic variables, i.e. points of the linguistic system which enable speakers to say the same thing in different ways, with these variants being *"identical in reference or truth value, but opposed in their social [...] significance"* [128]. Such variables have been described in many languages: variable pronunciation of -ing as [in] instead of [iŋ] in English (*playing* pronounced *playin'*); optional realization of the first part of the French negation (*je (ne) fume pas*, "I do not smoke"); optional realization of the plural ending of verb in Brazilian Portuguese (eles disse(ram), "they said"). For decades, sociolinguistic studies have showed that hearing certain variants triggers social stereotypes [30]. The so-called standard variants (*e.g.* [iŋ], realization of negative ne and plural -ram) are associated with social prestige, high education, professional ambition and effectiveness. They are more often produced in more formal situation. Non-standard variants are linked to social skills, solidarity and loyalty towards the local group, and they are produced more frequently in less formal situation.

It is therefore reasonable to say that the sociolinguistic task can benefit from the rapid development of computational social science [132]: the similarity of the online communication and face-to-face interaction [87] ensures the validity of the comparison with previous works. In this context, the nascent field of computational sociolinguistics found the digital counterparts of the sociolinguistic patterns already observed in spoken interaction. Our work lies at the core of this burgeoning field. It constructs the largest dataset of French tweets enriched with census sociodemographic information existent to date to the best of our knowledge. From this dataset, we observed variation of two grammatical cues and an index of vocabulary size in users located in France. We study how the linguistic cues correlated with three features reflective of the socioeconomic status of the users, their most representative location and their daily periods of activity on Twitter. We also observed whether connected people are more linguistically alike than disconnected ones. Multivariate analysis shows strong correlations between linguistic cues and socioeconomic status as well as a broad spatial pattern never observed before, with more standard language variants and lexical

diversity in the southern part of the country. Moreover, we found an unexpected daily cyclic evolution of the frequency of standard variants. Further analysis revealed that the observed cycle arose from the ever changing average economic status of the population of users present in Twitter through the day. Finally, we were able to establish that linguistic similarity between connected people does arises partially but not uniquely due to status homophily (users with similar socioeconomic status are linguistically similar and tend to connect). Its emergence is also due to other effects potentially including other types of homophilic correlations or influence disseminated over links of the social network. Beyond, we verify the presence of status homophily in the Twitter social network our results may inform novel methods to infer socioeconomic status of people from the way they use language. Furthermore, our work, rooted within the web content analysis line of research [95], extends the usual focus on aggregated textual features (like document frequency metrics or embedding methods) to specific linguistic markers, thus enabling sociolinguistics knowledge to inform the data collection process.

2.2 Related Work

For decades, sociolinguistic studies have repeatedly shown that speakers vary the way they talk depending on several factors. These studies have usually been limited to the analysis of small scale datasets, often obtained by surveying a set of individuals, or by direct observation after placing them in a controlled experimental setting. In spite of the volume of data collected generally, these studies have consistently shown the link between linguistic variation and social factors [18, 129].

Recently, the advent of social media and publicly available communication platforms has opened up a new gate to access individual information at a massive scale. Among all available social platforms, Twitter has been regarded as the choice by default, namely thanks to the intrinsic nature of communications taking place through it and the existence of data providers that are able to supply researchers with the volume of data they require. Work previously done on demographic variation is now relying increasingly on corpora from this social media platform as evidenced by the myriad of results showing that this resource reflects not only morpholexical variation of spoken language but also geographical [54, 172].

Although the value of this kind of platform for linguistic analysis has been more than proven, the question remains on how previous sociolinguistic results scale up to the sheer amount of data within reach and how can the latter enrich the former. To do so, numerous studies have focused on enhancing the data emanating from Twitter itself. Indeed, one of the core limitations of Twitter is the lack of reliable

sociodemographic information about the sampled users as usually data fields such as user-entered profile locations, gender or age differ from reality. This in turn implies that user-generated profile content cannot be used as a useful proxy for the sociodemographic information [73].

Many studies have overcome this limitation by taking advantage of the geolocation feature allowing Twitter users to include in their posts the location from which they were tweeted. Based on this metadata, studies have been able to assign home location to geolocated users with varying degrees of accuracy [4]. Subsequent work has also been devoted to assigning to each user some indicator that might characterize their socioeconomic status based on their estimated home location. These indicators are generally extracted from other datasets used to complete the Twitter one, namely census data [53, 54, 149] or real estate online services as Zillow.com [53]. Despite the relative success of these methods, their common limitation is to provide observations and predictions based on a carefully hand-picked small set of users, letting alone the problem of socioeconomic status inference on larger and more heterogeneous populations. Our work stands out from this well-established line of research by expanding the definition of socioeconomic status to include several demographic features as well as by pinpointing potential home location to individual users with an unprecedented accuracy. Identifying socioeconomic status and the network effects of homophily[195] is an open question [55]. However, recent results already showed that status homophily, i.e. the tendency of people of similar socioeconomic status are better connected among themselves, induce structural correlations which are pivotal to understand the stratified structure of society [135]. While we verify the presence of status homophily in the Twitter social network, we detect further sociolinguistic correlations between language, location, socioeconomic status, and time, which may inform novel methods to infer socioeconomic status for a broader set of people using common information available on Twitter.

2.3 Data Description

One of the main achievements of our study was the construction of a combined dataset for the analysis of sociolinguistic variables as a function of socioeconomic status, geographic location, time, and the social network. As follows, we introduce the two aforementioned independent datasets and how they were combined. We also present a brief cross-correlation analysis to ground the validity of our combined dataset for the rest of the study. In what follows, it should also be noted that regression analysis was performed via linear regression as implemented in the Scikit Learn Toolkit while data preprocessing and network study were performed using respectively pandas [156] and NetworkX [198] Python libraries.

2.3.1 Twitter dataset: sociolinguistic features

Our first dataset consists of a large data corpus collected from the online news and social networking service, Twitter. On it, users can post and interact with messages, "tweets", restricted to 140 characters. Tweets may come with several types of metadata including information about the author's profile, the detected language, where and when the tweet was posted, etc. Specifically, we recorded 170 million tweets written in French, posted by 2.5 million users in the timezones GMT and GMT+1 over three years (between July 2014 to May 2017). These tweets were obtained via the Twitter powertrack API feeds provided by Datasift and Gnip with an access rate varying between $15 - 25\%$ [1].

Linguistic data: To obtain meaningful linguistic data we preprocessed the incoming tweet stream in several ways. As our central question here deals with the variability of the language, repeated tweets do not bring any additional information to our study. Therefore, as an initial filtering step, we decided to remove retweets. Next, in order to facilitate the detection of the selected linguistic markers we removed any URLs, emoticons, mentions of other users (denoted by the `@` symbol) and hashtags (denoted by the `#` symbol) from each tweet. These expressions were not considered to be semantically meaningful and their filtering allowed to further increase the speed and accuracy of our linguistic detection methods when run across the data. In addition we completed a last step of textual preprocessing by down-casing and stripping the punctuation out of the tweets body. POS-taggers such as MElt [49] were also tested but they provided no significant improvement in the detection of the linguistic markers.

Network data: We used the collected tweets in another way to infer social relationships between users. Tweet messages may be direct interactions between users, who mention each other in the text by using the `@` symbol (`@username`). When one user u, mentions another user v, user v will see the tweet posted by user u directly in his / her feed and may tweet back. In our work we took direct mentions as proxies of social interactions and used them to identify social ties between pairs of users. Opposite to the follower network, reflecting passive information exposure and less social involvement, the mutual mention network has been shown [100] to capture better the underlying social structure between users. We thus use this network definition in our work as links are a greater proxy for social interactions.

In our definition we assumed a tie between users if they mutually mentioned each

1. In order to uphold the strict privacy laws in France as well as the agreement signed with our data provider GNIP, full disclosure of the original dataset is not possible. Data collection and preprocessing pipelines could however be released upon request.

other at least once during the observation period. People who reciprocally mentioned each other express some mutual interest, which may be a stronger reflection of real social relationships as compared to the non-mutual cases [94]. This constraint reduced the egocentric social network considerably leading to a directed structure of 508, 975 users and 4, 029, 862 links that we considered being undirected in what follows.

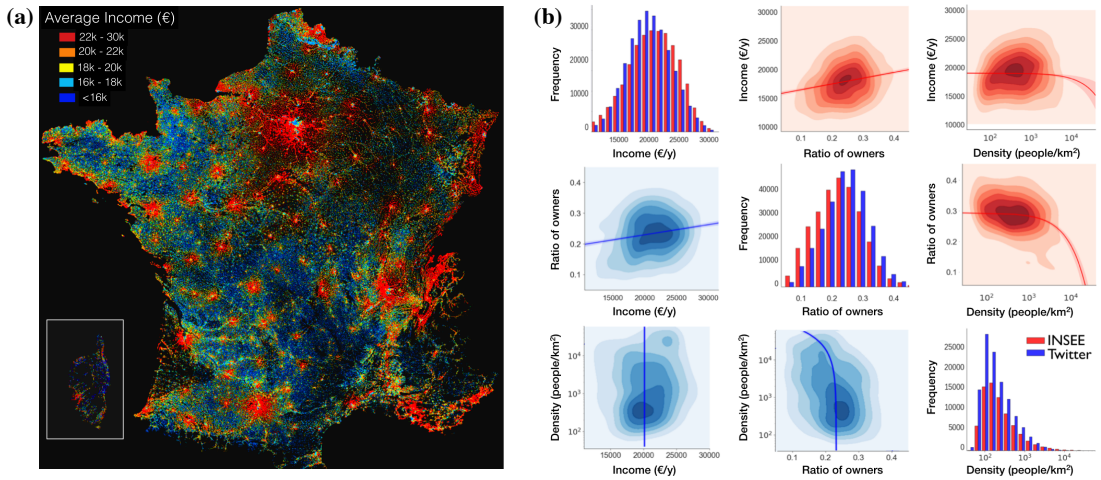


Figure 2-1: Distributions and correlations of socioeconomic indicators. (a) Spatial distribution of average income in France with $200m \times 200m$ resolution. (b) Distribution of socioeconomic indicators (in the diag.) and their pairwise correlations measured in the INSEE (upper diag. panels) and Twitter geotagged (lower diag. panels) datasets. Contour plots assign the equidensity lines of the scatter plots, while solid lines are the corresponding linear regression values. Population density in log.

Geolocated data: About 2% of tweets included in our dataset contained some location information regarding either the tweet author's self-provided position or the place from which the tweet was posted. These pieces of information appeared as the combination of self reported locations or usual places tagged with GPS coordinates at different geographic resolution. We considered only tweets which contained the exact GPS coordinates with resolution of ~ 3 meters of the location where the actual tweet was posted. This actually means that we excluded tweets where the user assigned a place name such as "Paris" or "France" to the location field, which are by default associated to the geographical center of the tagged areas. Practically, we discarded coordinates that appeared more than 500 times throughout the whole GPS-tagged data, assuming that there is no such 3×3 meter rectangle in the country where 500 users could appear and tweet by chance. After this selection procedure we rounded up each tweet location to a 100 meter precision.

To obtain a unique representative location of each user, we extracted the sequence of all declared locations from their geolocated tweets. Using this set of locations we

selected the most frequent to be the representative one, and we took it as a proxy for the user's home location. Further we limited our users to ones located throughout the French territory thus not considering others tweeting from places outside the country. This selection method provided us with 110, 369 geolocated users who are either detected as French speakers or assigned to be such by Twitter and all associated to specific 'home' GPS coordinates in France. To verify the spatial distribution of the selected population, we further assessed the correlations between the true population distributions (obtained from census data [104]) at different administrative level and the geolocated user distribution aggregated correspondingly. More precisely, we computed the R^2 coefficient of variation between the inferred and official population distributions (a) at the level of 22 regions [2]. Correlations at this level induced a high coefficient of $R^2 \simeq 0.89$ ($p < 10^{-2}$); (b) At the arrondissement level with 322 administrative units and coefficient $R^2 \simeq 0.87$ ($p < 10^{-2}$); and (c) at the canton level with 4055 units with a coefficient $R \simeq 0.16$ ($p < 10^{-2}$). Note that the relatively small coefficient at this level is due to the interplay of the sparsity of the inferred data and the fine grained spatial resolution of cantons. All in all, we can conclude that our sample is highly representative in terms of spatial population distribution, which at the same time validate our selection method despite the potential inherent biases induced by the method taking the most frequented GPS coordinates as the user's home location.

2.3.2 INSEE dataset: socioeconomic features

The second dataset we used was released in December 2016 by the National Institute of Statistics and Economic Studies (INSEE) of France. This data corpus [104] contains a set of sociodemographic aggregated indicators, estimated from the 2010 tax return in France, for each 4 hectare ($200m \times 200m$) square patch across the whole French territory. Using these indicators, one can estimate the distribution of the average socioeconomic status (SES) of people with high spatial resolution. In this study, we concentrated on three indicators for each patch i , which we took to be good proxies of the socioeconomic status of the people living within them. These were the S^i_{inc} average yearly income per capita (in euros), the S^i_{own} fraction of owners (not renters) of real estate, and the S^i_{den} density of population defined respectively as

$$: S^i_{\text{inc}} = \frac{S^i_{hh}}{N^i_{hh}}, \quad S^i_{\text{own}} = \frac{N^i_{\text{own}}}{N^i}, \quad \text{and} \quad S^i_{\text{den}} = \frac{N^i}{(200m)^2}. \tag{2.1}$$

Here S^i_{hh} and N^i_{hh} assign respectively the cumulative income and total number of inhabitants of patch i, while N^i_{own} and N^i are respectively the number of real estate

2. Note that since 2016 France law determines 13 metropolitan regions, however the available data shared by INSEE [104] contained information about the earlier administrative structure containing 22 regions.

owners and the number of individuals living in patch i. As an illustration we show the spatial distribution of S^i_{inc} average income over the country in Fig.2-1a.

In order to uphold current privacy laws and due to the highly sensitive nature of the disclosed data, some statistical pretreatments were applied to the data by INSEE before its public release. More precisely, neighboring patches with less than 11 households were merged together, while some of the sociodemographic indicators were winsorized. This set of treatments induced an inherent bias responsible for the deviation of the distribution of some of the socioeconomic indicators. These quantities were expected to be determined by the Pareto principle, thus reflecting the high level of socioeconomic imbalances present within the population. Instead, as shown in Fig.2-1b [diagonal panels], distributions of the derived socioeconomic indicators (in blue) appeared somewhat more symmetric than expected. This doesn't hold though for $P(S^i_{\text{den}})$ (shown on a log-log scale in the lowest right panel of Fig.2-1b), which emerged with a broad tail similar to an expected power-law Pareto distribution. In addition, although the patches are relatively small ($200m \times 200m$), the socioeconomic status of people living may have some local variance, what we cannot consider here. Nevertheless, all things considered, this dataset and the derived socioeconomic indicators yield the most fine-grained description, allowed by national law, about the population of France over its whole territory.

Despite the inherent biases of the selected socioeconomic indicators, in general we found weak but significant pairwise correlations between these three variables as shown in the upper diagonal panels in Fig.2-1b (in red), with values in Table 2.1. We observed that while S^i_{inc} income and S^i_{own} owner ratio are positively correlated ($R = 0.24$, $p < 10^{-2}$), and the S^i_{own} and S^i_{den} population density are negatively correlated ($R = -0.23$, $p < 10^{-2}$), S^i_{inc} and S^i_{den} appeared to be very weakly correlated ($R = -0.07$, $p < 10^{-2}$). This nevertheless suggested that high average income, high owner ratio, and low population density are consistently indicative of high socioeconomic status in the dataset.

Table 2.1: Pearson correlations and p-values measured between SES indicators in the INSEE and Twitter datasets.

	$S^i_{\text{inc}} \sim S^i_{\text{own}}$	$S^i_{\text{inc}} \sim S^i_{\text{den}}$	$S^i_{\text{own}} \sim S^i_{\text{den}}$
INSEE	0.24 ($p < 10^{-2}$)	-0.07 ($p < 10^{-2}$)	-0.23 ($p < 10^{-2}$)
Twitter	0.19 ($p < 10^{-2}$)	0.00 ($p > 10^{-2}$)	-0.22 ($p < 10^{-2}$)

2.3.3 Combined dataset: individual socioeconomic features

Data collected from Twitter provides a large variety of information about several users including their tweets, which disclose their interests, vocabulary, and linguistic patterns; their direct mentions from which their social interactions can be inferred; and the sequence of their locations, which can be used to infer their representative location. However, no information is directly available regarding their socioeconomic status, which can be pivotal to understand the dynamics and structure of their personal linguistic patterns.

To overcome this limitation we combined our Twitter data with the socioeconomic maps of INSEE by assigning each geolocated Twitter user to a patch closest to their estimated home location (within 1 km). This way we obtained for all $110,369$ geolocated users their dynamical linguistic data, their egocentric social network as well as a set of SES indicators.

Such a dataset associating language with socioeconomic status and social network throughout the French metropolitan territory is unique to our knowledge and provides unrivaled opportunities to verify sociolinguistic patterns observed over a long period on a small-scale, but never established in such a large population.

To verify whether the geolocated Twitter users yet provide a representative sample of the whole population we compared the distribution and correlations of the their SES indicators to the population measures. Results are shown in Fig.2-1b diagonal (red distributions) and lower diagonal panels (in blue) with correlation coefficients and p-values summarized in Table.2.1. Even if we observed some discrepancy between the corresponding distributions and somewhat weaker correlations between the SES indicators, we found the same significant correlation trends (with the exception of the pair density / income) as the ones seen when studying the whole population, assuring us that each indicator correctly reflected the SES of individuals.

2.3.4 Socioeconomic Class Definition

In order to assign users to a particular socioeconomic class, we took the geolocated Twitter users in France and partitioned them into nine socioeconomic classes using their inferred income S^u_{inc}. Partitioning was done first by sorting users by their S^u_{inc} income to calculate their $C(S^u_{\text{inc}})$ cumulative income distribution function. We defined socioeconomic classes by segmenting $C(S^u_{\text{inc}})$ such that the sum of income is the same for each classes (for an illustration of our method see Fig.2-6a).

2.4 Linguistic variables

We identified the following three linguistic markers to study across users from different socioeconomic backgrounds: Correlation with SES has been evidenced for all of them. The optional deletion of negation is typical of spoken French, whereas the omission of the mute letters marking the plural in the nominal phrase is a variable cue of French writing. The third linguistic variable is a global measure of the lexical diversity of the Twitter users. We present them here in greater detail.

2.4.1 Standard usage of negation

The basic form of negation in French includes two negative particles: *ne* (no) before the verb and another particle after the verb that conveys more accurate meaning: *pas* (not), *jamais* (never), *personne* (no one), *rien* (nothing), etc. Due to this double construction, the first part of the negation (*ne*) is optional in spoken French, but it is obligatory in standard writing.

Sociolinguistic studies have previously observed the realization of *ne* in corpora of recorded everyday spoken interactions. Although all the studies do not converge, a general trend is that *ne* realization is more frequent in speakers with higher socioeconomic status than in speakers with lower status [8, 16]. We built upon this research to set out to detect both negation variants in the tweets using regular expressions. [3] We are namely interested in the rate of usage of the standard negation (featuring both negative particles) across users:

$$L^u_{\text{cn}} = \frac{n^u_{\text{cn}}}{n^u_{\text{cn}} + n^u_{\text{incn}}} \quad \text{and} \quad \overline{L}^i_{\text{cn}} = \frac{\sum_{u \in i} L^u_{\text{cn}}}{N_i}, \tag{2.2}$$

where n^u_{cn} and n^u_{incn} assign the number of correct negation and incorrect number of negation of user u, thus L^u_{cn} defines the rate of correct negation of a users and $\overline{L}^i_{\text{cn}}$ its average over a selected i group (like people living in a given place) of N_i users.

2.4.2 Standard usage of plural ending of written words

In written French, adjectives and nouns are marked as being plural by generally adding the letters *s* or *x* at the end of the word. Because these endings are mute (without counterpart in spoken French), their omission is the most frequent spelling error in adults [151]. Moreover, studies showed correlations between standard spelling and social status of the writers, in preteens, teens and adults [23, 151, 221]. We then

3. Negation:`\\b(pas|pa|aps|jamais|ni|personne|rien|ri1|r1|aucun|aucune)\\b`
Standard Negation:. * `\\b(ne|n')\\b`. * `\$`

set to estimate the use of standard plural across users:

$$L^u_{\text{cp}} = \frac{n^u_{\text{cp}}}{n^u_{\text{cp}} + n^u_{\text{incp}}} \quad \text{and} \quad \overline{L}^i_{\text{cp}} = \frac{\sum_{u \in i} L^u_{\text{cp}}}{N_i} \tag{2.3}$$

where the notation follows as before (cp stands for correct plural and incp stands for incorrect plural).

2.4.3 Normalized vocabulary set size

A positive relationship between an adult's lexical diversity level and his or her socioeconomic status has been evidenced in the field of language acquisition. Specifically, converging results showed that the growth of child lexicon depends on the lexical diversity in the speech of the caretakers, which in turn is related to their socioeconomic status and their educational level [91, 101]. We thus proceeded to study the following metric:

$$L^u_{\text{vs}} = \frac{N^u_{\text{vs}}}{N^u_{tw}} \quad \text{and} \quad \overline{L}^i_{\text{vs}} = \frac{\sum_{u \in i} N^u_{\text{vs}}}{N_i}, \tag{2.4}$$

where $N_v s^u$ assigns the total number of unique words used by user u who tweeted N^u_{tw} times during the observation period. As such L^u_{vs} gives the normalized vocabulary set size of a user u, while $\overline{L}^i_{\text{vs}}$ defines its average for a population i.

2.5 Results

By measuring the defined linguistic variables in the Twitter timeline of users we were finally set to address the core questions of our study, which dealt with linguistic variation. More precisely, we asked whether the language variants used online depend on the socioeconomic status of the users, on the location or time of usage, and on ones social network. To answer these questions we present here a multidimensional correlation study on a large set of Twitter geolocated users, to which we assigned a representative location, three SES indicators, and a set of meaningful social ties based on the collection of their tweets.

2.5.1 Socioeconomic variation

The socioeconomic status of a person is arguably correlated with education level, income, habitual location, or even with ethnicity and political orientation and may strongly determine to some extent patterns of individual language usage. Such dependencies have been theoretically proposed before [129], but have rarely been inspected at this scale yet. The use of our previously described datasets enabled us to do so via the measuring of correlations between the inferred SES indicators of Twitter

users and the use of the previously described linguistic markers which we expand upon in this section. To compute and visualize these correlations we defined linear

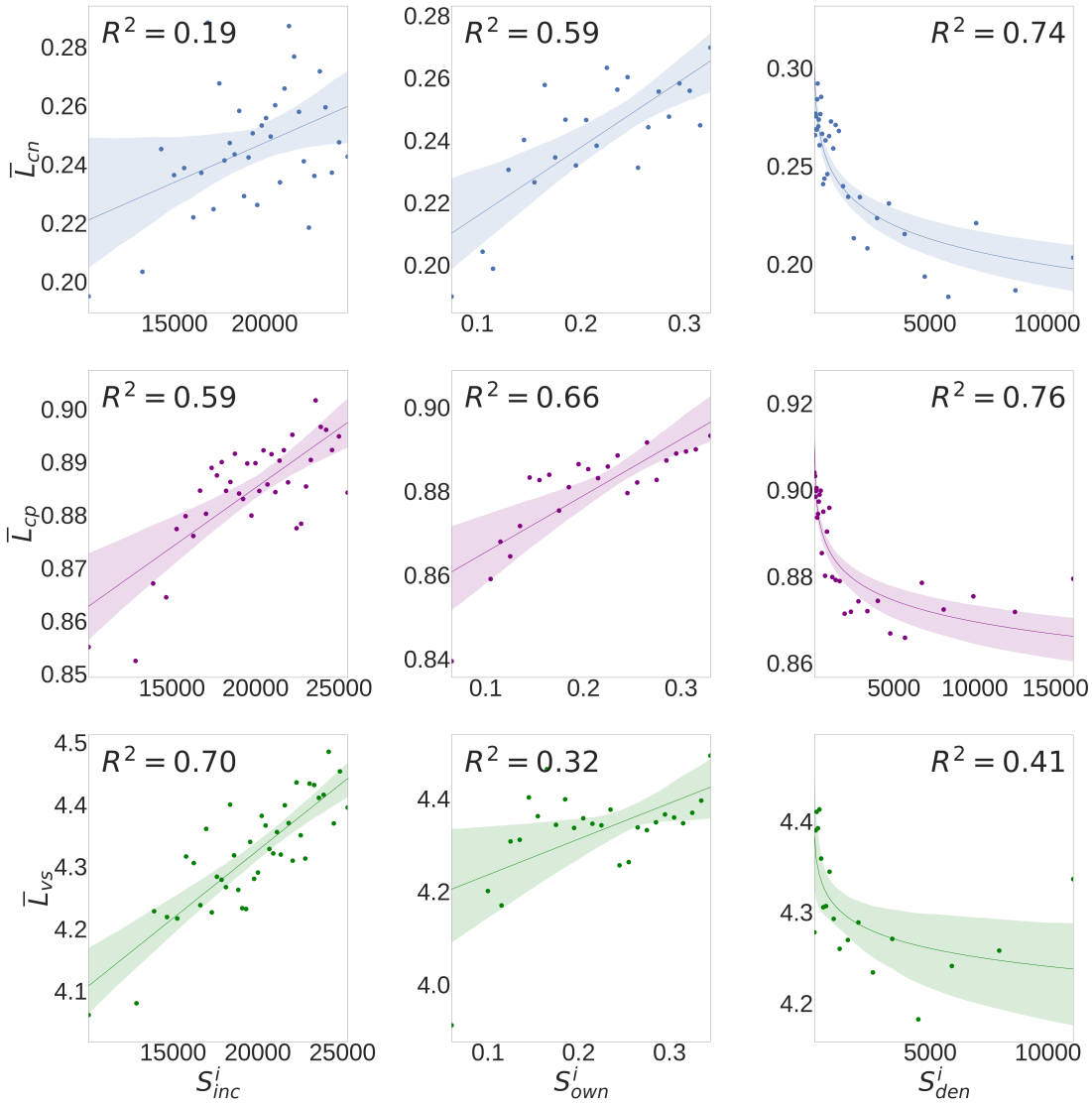


Figure 2-2: Pairwise correlations between three SES indicators and three linguistic markers. Columns correspond to SES indicators (resp. $S^i_{\rm inc}$, $S^i_{\rm own}$, $S^i_{\rm den}$), while rows correspond to linguistic variables (resp. $\overline{L}_{\rm cn}$, $\overline{L}_{\rm cp}$ and $\overline{L}_{\rm vs}$). On each plot colored symbols are binned data values and a linear regression curve are shown together with the 95 percentile confidence interval and R^2 values.

bins (in numbers varying from 20 to 50) for the socioeconomic indicators and computed the average of the given linguistic variables for people falling within the given bin. These binned values (shown as symbols in Fig.2-2) were used to compute linear regression curves and the corresponding confidence intervals (see Fig.2-2). An additional transformation was applied to the SES indicator describing population density, which was broadly distributed (as discussed in Section 2.3.2 and Fig.2-1b), thus, for

the regression process, the logarithm of its values were considered. To quantify pairwise correlations we computed the R^2 coefficient of determination values in each case.

In Fig.2-2 we show the correlation plots of all nine pairs of SES indicators and

Table 2.2: The R^2 coefficient of determination and the corresponding p-values computed for the pairwise correlations of SES indicators and linguistic variables.

	S^i_{inc}	S^i_{own}	S^i_{den}
$\overline{L}_{\text{cn}}$	0.19 ($p < 10^{-2}$)	0.59 ($p < 10^{-2}$)	0.74 ($p < 10^{-2}$)
$\overline{L}_{\text{cp}}$	0.59 ($p < 10^{-2}$)	0.66 ($p < 10^{-2}$)	0.76 ($p < 10^{-2}$)
$\overline{L}_{\text{vs}}$	0.70 ($p < 10^{-2}$)	0.32 ($p < 10^{-2}$)	0.41 ($p < 10^{-2}$)

linguistic variables together with the linear regression curves, the corresponding R^2 values and the 95 percentile confidence intervals (note that all values are also in Table 2.2). These results show that correlations between socioeconomic indicators and linguistic variables actually exist. Furthermore, these correlation trends suggest that people with lower SES may use more non-standard expressions (higher rates of incorrect negation and plural forms) have a smaller vocabulary set size than people with higher SES. Note that, although the observed variation of linguistic variables were limited, all the correlations were statistically significant ($p < 10^{-2}$) with considerably high R^2 values ranging from 0.19 (between $\overline{L}_{\text{cn}} \sim S_{\text{inc}}$) to 0.76 (between $\overline{L}_{\text{cp}} \sim S_{\text{den}}$). For the rates of standard negation and plural terms the population density appeared to be the most determinant indicator with $R^2 = 0.74$ (and 0.76 respectively), while for the vocabulary set size the average income provided the highest correlation (with $R^2 = 0.7$).

One must also acknowledge that while these correlations exhibit high values consistently across linguistic and socioeconomic indicators, they only hold meaning at the population level at which the binning was performed. When the data is considered at the user level, the variability of individual language usage hinders the observation of the aforementioned correlation values.

2.5.2 Spatial variation

Next we will cover the spatial variation of linguistic variables. Although officially a standard language is used over the whole country, geographic variations of the former may exist due to several reasons [123, 228]. For instance, regional variability resulting from remnants of local languages that have disappeared, uneven spatial distribution of socioeconomic potentials, or influence spreading from neighboring countries might play a part in this process. For the observation of such variability, by using their representative locations, we assigned each user to a department of France. We then

computed the $\overline{L}^i_{\text{cn}}$ (resp. $\overline{L}^i_{\text{cp}}$) average rates of standard negation (resp. plural agreement) and the $\overline{L}^i_{\text{vs}}$ average vocabulary set size for each "département" i in the country (administrative division of France – There are 97 départements).

Results shown in Fig.2-3a-c revealed some surprising patterns, which appeared

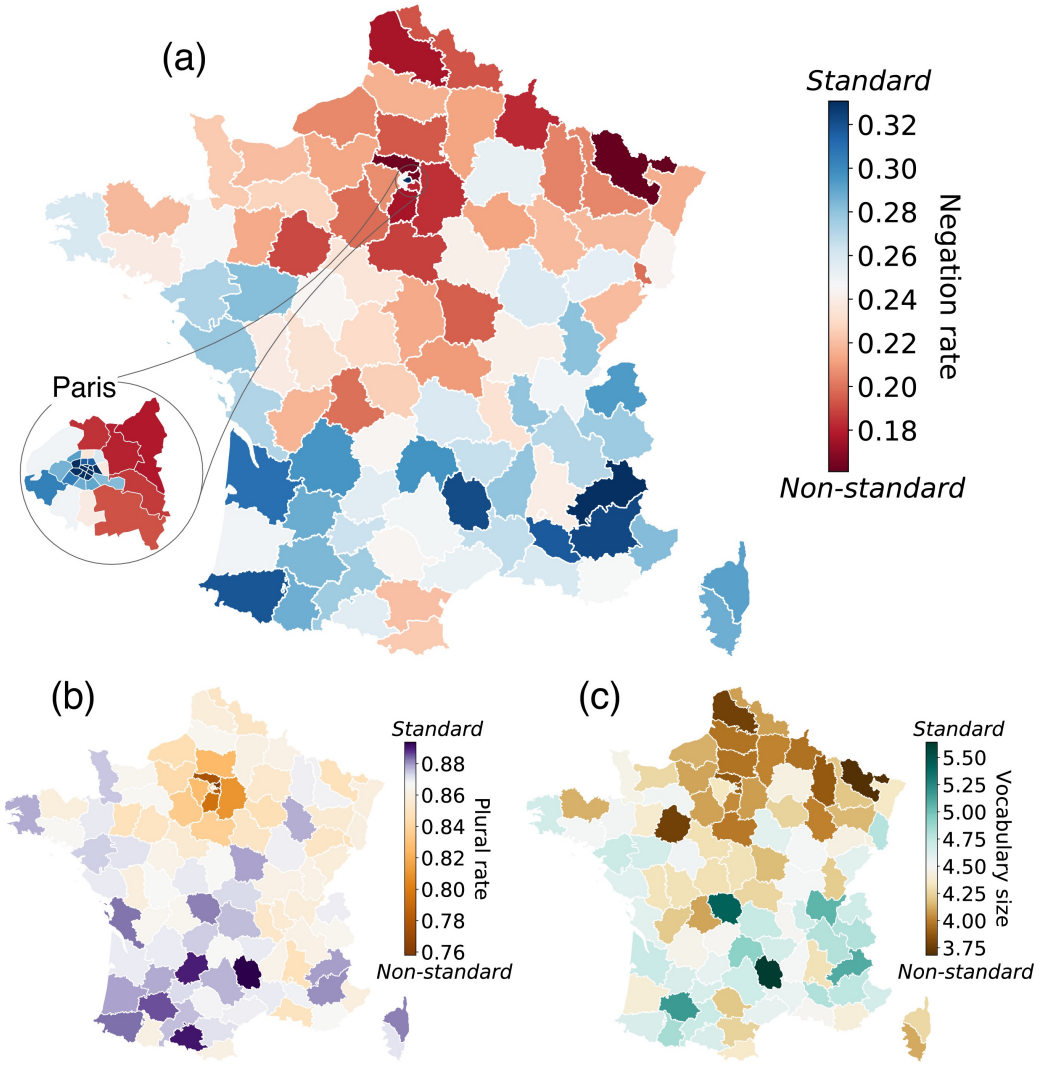


Figure 2-3: Geographical variability of linguistic markers in France. (a) Variability of the rate of correct negation. Inset focuses on larger Paris. (b) Variability of the rate of correct plural terms. (c) Variability of the average vocabulary size set. Each plot depicts variability on the department level except the inset of (a) which is on the "arrondissements" level.

to be consistent for each linguistic variable. By considering latitudinal variability it appeared that, overall, people living in the northern part of the country used a less standard language, i.e., negated and pluralized less standardly, and used a smaller number of words. On the other hand, people from the South used a language which

is somewhat closer to the standard (in terms of the aforementioned linguistic markers) and a more diverse vocabulary. The most notable exception is Paris, where in the city center people used more standard language, while the contrary is true for the suburbs. This observation, better shown in Fig.2-3a inset, can be explained by the large differences in average socioeconomic status between districts. Such segregation is known to divide the Eastern and Western sides of suburban Paris, and in turn to induce apparent geographic patterns of standard language usage. We found less evident longitudinal dependencies of the observed variables. Although each variable shows a somewhat diagonal trend, the most evident longitudinal dependency appeared for the average rate of standard pluralization (see Fig.2-3b), where users from the Eastern side of the country used the language in less standard ways. Note that we also performed a multivariate regression analysis (not shown here), using the linguistic markers as target and considering as factors both location (in terms of latitude and longitude) as and income as proxy of socioeconomic status. It showed that while location is a strong global determinant of language variability, socioeconomic variability may still be significant locally to determine standard language usage (just as we demonstrated in the case of Paris).

2.5.3 Temporal variation

Another potentially important factor determining language variability is the time of day when users are active in Twitter which we now proceed to study[78, 122]. The temporal variability of standard language usage can be measured for a dynamical quantity like the $L_{\rm cn}(t)$ rate of correct negation. To observe its periodic variability (with a ΔT period of one week) over an observation period of T (in our case 734 days), we computed

$$\overline{L}^{\Lambda}_{\rm cn}(t) = \frac{\Delta T}{|\Lambda| T} \sum_{u \in \Lambda} \sum_{k=0}^{\lfloor T/\Delta T \rfloor} L^{u}_{\rm cn}(t + k\Delta T), \tag{2.5}$$

in a population Λ of size $|\Lambda|$ with a time resolution of one hour. This quantity reflects the average standard negation rate in an hour over the week in the population Λ. Note that an equivalent $\overline{L}^{\Lambda}_{\rm cp}(t)$ measure can be defined for the rate of standard plural terms, but not for the vocabulary set size as it is a static variable.

In Fig. 2-4a and b we show the temporal variability of $\overline{L}^{\Lambda}_{\rm cn}(t)$ and $\overline{L}^{\Lambda}_{\rm cp}(t)$ (respectively) computed for the whole Twitter user set ($\Gamma = all$, solid line) and for geolocated users ($\Gamma = geo$, dashed lines). Not surprisingly, these two curves were strongly correlated as indicated by the high Pearson correlation coefficients summarized in the last column of Table 2.3 which, again, assured us that our geolocated sample of Twitter users was representative of the whole set of users. At the same time, the temporal

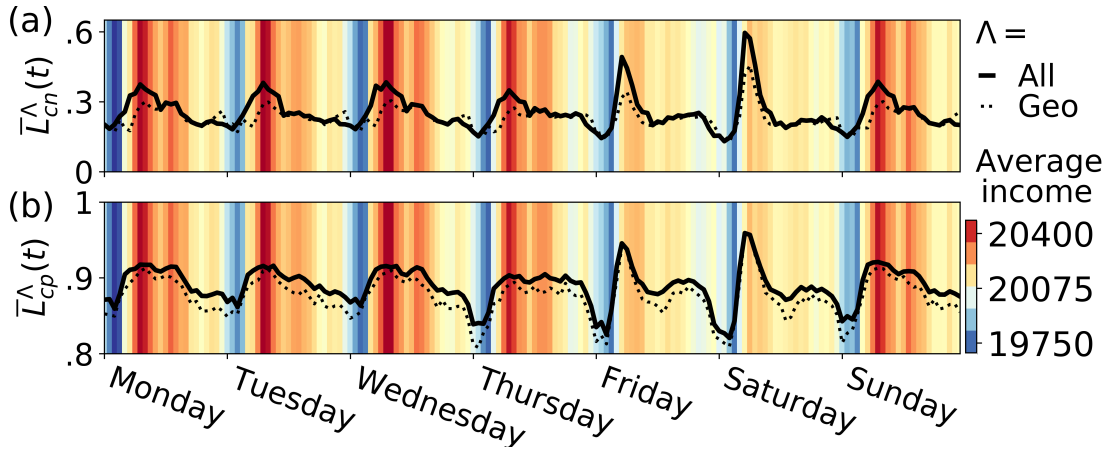


Figure 2-4: Temporal variability of (a) $\overline{L}^{\Lambda}_{\text{cn}}(t)$ (resp. (b) $\overline{L}^{\Lambda}_{\text{cp}}(t)$) average rate of correct negation (resp. plural terms) over a week with one hour resolution. Rates were computed for $\Lambda = all$ (solid line) and $\Lambda = geo$located Twitter users. Colors indicates the temporal variability of the average income of geolocated population active in a given hour.

variability of these curves suggested that people tweeting during the day used a more standard language than those users who are more active during the night. However, after measuring the average income of active users in a given hour over a week, we obtained an even more sophisticated picture. It turned out that people active during the day have higher average income (warmer colors in Fig. 2-4) than people active during the night (colder colors in Fig. 2-4). Thus the variability of standard language patterns was largely explained by the changing overall composition of active Twitter users during different times of day and the positive correlation between socioeconomic status and the usage of higher linguistic standards (that we have seen earlier). This explanation was supported by the high coefficients (summarized in Table 2.3), which were indicative of strong and significant correlations between the temporal variability of average linguistic variables and average income of the active population on Twitter.

Table 2.3: Pearson correlations and p-values of pairwise correlations of time varying $S_{\text{inc}}(t)$ average income with $\overline{L}^{\Lambda}_{\text{cn}}(t)$ and $\overline{L}^{\Lambda}_{\text{cp}}(t)$ average linguistic variables; and between average linguistic variables of $\Lambda = all$ and $\Lambda =$ geo-localized users.

	$\overline{L}^{all}_{*}(t) \sim S_{\text{inc}}(t)$	$\overline{L}^{geo}_{*}(t) \sim S_{\text{inc}}(t)$	$\overline{L}^{geo}_{*}(t) \sim \overline{L}^{all}_{*}(t)$
$* = \text{cn}$	0.5915 ($p < 10^{-2}$)	0.622 ($p < 10^{-2}$)	0.805 ($p < 10^{-2}$)
$* = \text{cp}$	0.7027 ($p < 10^{-2}$)	0.665 ($p < 10^{-2}$)	0.98021 ($p < 10^{-2}$)

2.5.4 Network variation

Finally we sought to understand the effect of the social network on the variability of linguistic patterns which we now expand upon. People in a social structure can be connected due to several reasons. Link creation mechanisms like focal or cyclic closure [125, 131], or preferential attachment together with the effects of homophily [157] are all potentially driving the creation of social ties and communities, and the emergence of community rich complex structure within social networks. In terms of homophily, one can identify several individual characteristics like age, gender, common interest or political opinion, etc., that might increase the likelihood of creating relationships between disconnected but similar people, who in turn influence each other and become even more similar. Status homophily between people of similar socioeconomic status has been shown to be important [135] in determining the creation of social ties and to explain the stratified structure of society. By using our combined datasets, we aim here to identify the effects of status homophily and to distinguish them from other homophilic correlations and the effects of social influence inducing similarities among already connected people. To do so, we constructed a social network by considering mutual mention links between these users (as introduced in Section 3.3).

Status homophily in social networks appears as an increased tendency for people from similar socioeconomic classes to be connected. This correlation can be identified by comparing likelihood of connectedness in the empirical network to a random network, which conserves all network properties except structural correlations. To do so, we took each (s_i, s_j) pair of the nine SES class in the Twitter network and counted the number of links $|E(s_i, s_j)|$ connecting people in classes s_i and s_j. As a reference system, we computed averages over 100 corresponding configuration model network structures [169].

To signalize the effects of status homophily, we took the ratio $|E(s_i, s_j)|/|E_{rand}(s_i, s_j)|$ of the two matrices (shown in Fig.2-6b). The diagonal component in Fig.2-6b with values larger than 1 showed that users of the same or similar socioeconomic class were better connected in the original structure than by chance, while the contrary was true for users from classes far apart (see blue off-diagonal components). To verify the statistical significance of this finding, we performed a χ^2-test, which showed that the distribution of links in the original matrix was significantly different from the one of the average randomized matrix ($p < 10^{-5}$). This observation verified status homophily present in the Twitter mention network. In order to measure linguistic similarities between a pair of users u and v, we simply computed the $|L^u_* - L^v_*|$ absolute difference of their corresponding individual linguistic variable $* \in \{\mathrm{cn}, \mathrm{cp}, vs\}$. This measure appeared with a minimum of 0 and associated smaller values to more similar

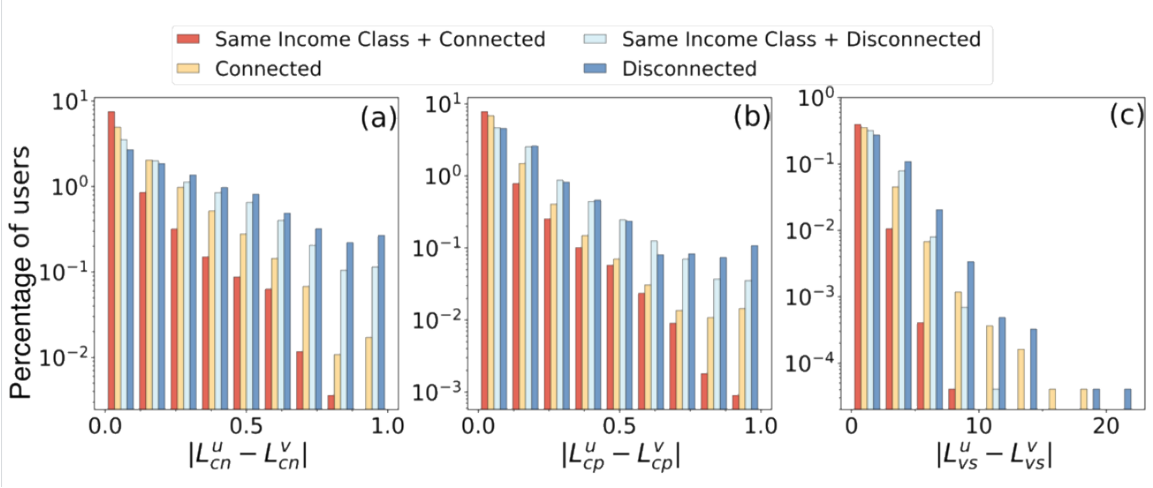


Figure 2-5: Distribution of the $|L^u_* - L^v_*|$ absolute difference of linguistic variables $* \in \{\text{cn}, \text{cp}, vs\}$ (resp. panels (a), (b), and (c)) of user pairs who were connected and from the same socioeconomic group (red), connected (yellow), disconnected and from the same socioeconomic group (light blue), disconnected pairs of randomly selected users (blue).

pairs of users. To identify the effects of status homophily and the social network, we proceeded by computing the similarity distribution in four cases: for connected users from the same socioeconomic class; for disconnected randomly selected pairs of users from the same socioeconomic class; for connected users in the network; and randomly selected pairs of disconnected users in the network. Note that in each case the same number of user pairs were sampled from the network to obtain comparable averages. This number was naturally limited by the number of connected users in the smallest socioeconomic class, and were chosen to be $10,000$ in each cases. By comparing the distributions shown in Fig.2-5 we concluded that (a) connected users (red and yellow bars) were the most similar in terms of any linguistic marker. This similarity was even greater when the considered tie was connecting people from the same socioeconomic group; (b) network effects can be quantified by comparing the most similar connected (red bar) and disconnected (light blue bar) users from the same socioeconomic group. Since the similarity between disconnected users here is purely induced by status homophily, the difference of these two bars indicates additional effects that cannot be explained solely by status homophily. These additional similarities may rather be induced by other factors such as social influence, the physical proximity of users within a geographical area or other homophilic effects that were not accounted for. (c) Randomly selected pairs of users were more dissimilar than connected ones as they dominated the distributions for larger absolute difference values. We therefore concluded that both the effects of network and status homophily mattered in terms of linguistic similarity between users of this social media platform.

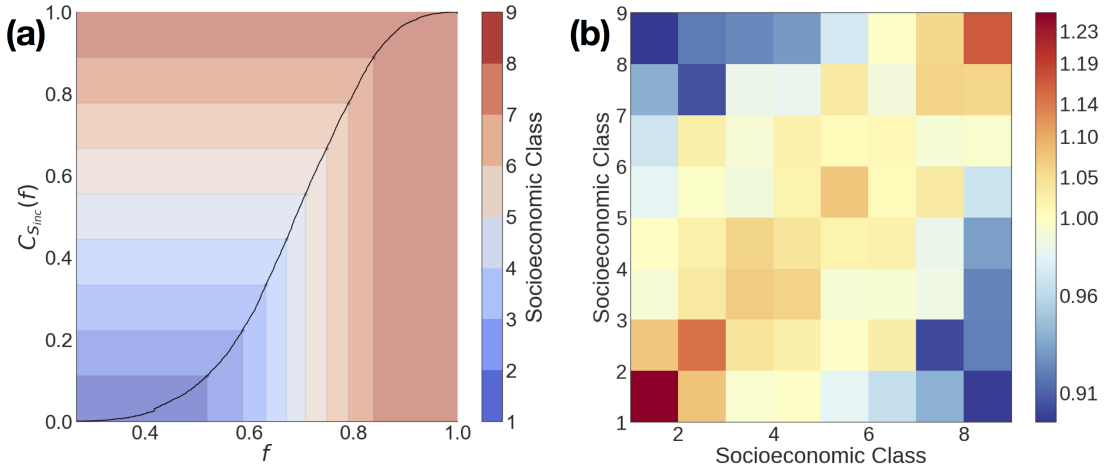


Figure 2-6: (a) Definition of socioeconomic classes by partitioning users into nine groups with the same cumulative annual income. (b) Structural correlations between SES groups depicted as matrix of the ratio $|E(s_i, s_j)|/|E_{rand}(s_i, s_j)|$ between the original and the average randomized mention network

2.6 Conclusions

The overall goal of this work was to explore the dependencies of linguistic variables on the socioeconomic status, location, time varying activity, and social network of users. To do so we constructed a combined dataset from a large Twitter data corpus, including geotagged posts and proxy social interactions of millions of users, as well as a detailed socioeconomic map describing average socioeconomic indicators with a high spatial resolution in France. The combination of these datasets provided us with a large set of Twitter users all assigned to their Twitter timeline over three years, their location, three individual socioeconomic indicators, and a set of meaningful social ties. Three linguistic variables extracted from individual Twitter timelines were then studied as a function of the former, namely, the rate of standard negation, the rate of plural agreement and the size of vocabulary set.

Via a detailed multidimensional correlation study we concluded that (a) socioeconomic indicators and linguistic variables are significantly correlated. i.e. people with higher socioeconomic status are more prone to use more standard variants of language and a larger vocabulary set, while people on the other end of the socioeconomic spectrum tend to use more non-standard terms and, on average, a smaller vocabulary set; (b) Spatial position was also found to be a key feature of standard language use as, overall, people from the North tended to use more non-standard terms and a smaller vocabulary set compared to people from the South; a more fine-grained analysis reveals that the spatial variability of language is determined to a greater extent locally by the socioeconomic status; (c) In terms of temporal activity, standard language was more likely to be used during the daytime while non-standard variants were predom-

inant during the night. We explained this temporal variability by the turnover of population with different socioeconomic status active during night and day; Finally (d) we showed that the social network and status homophily mattered in terms of linguistic similarity between peers, as connected users with the same socioeconomic status appeared to be the most similar, while disconnected people were found to be the most dissimilar in terms of their individual use of the previous linguistic markers.

Despite these findings, one has to acknowledge the multiple limitations affecting this work: First of all, although Twitter is a broadly adopted service in most technologically enabled societies, it commonly provides a biased sample in terms of age and socioeconomic status as older or poorer people may not have access to this technology. In addition, home locations inferred for lower activity users may induced some noise in our inference method. Nevertheless, we demonstrated that our selected Twitter users are quite representative in terms of spatial, temporal, and socioeconomic distributions once compared to census data. Other sources of bias include the "homogenization" performed by INSEE to ensure privacy rights are upheld as well as the proxies we devised to approximate users' home location and social network. Currently, a sample survey of our set of geolocated users is being conducted so as to bootstrap socioeconomic data to users and definitely validate our inference results. Nonetheless, this INSEE dataset provides still the most comprehensive available information on socioeconomic status over the whole country. For limiting such risk of bias, we analyzed the potential effect of the confounding variables on distribution and cross-correlations of SES indicators. Acknowledging possible limitations of this study, we consider it as a necessary first step in analyzing income through social media using datasets orders of magnitude larger than in previous research efforts.

Finally we would like to emphasize two scientific merits of this work. On one side, based on a very large sample, we confirm and clarify results from the field of sociolinguistics and we highlight new findings. We thus confirm clear correlations between the variable realization of the negative particle in French and three indices of socioeconomic status. This result challenges those among the sociolinguistic studies that do not find such correlation. Our data also suggested that the language used in the southern part of France is more standard. Understanding this pattern fosters further investigations within sociolinguistics. We finally established that the linguistic similarity of socially connected people is partially explained by status homophily but could be potentially induced by social influences passing through the network of links or other terms of homophilic correlations. Our results highlight the way ahead by showcasing how we might rely on linguistic information, available on Twitter and other online platforms, to infer socioeconomic status of individuals from their position in the network as well as the way they use their language. This question will therefore be addressed in the coming chapter.

Chapter 3

Language based Socioeconomic Status Inference

This work was originally published in the Proceedings of the Workshop on Social Computing [2] and later extended as a journal paper in Complexity [142] with myself as leading author

3.1 Introduction

The previous chapter compiled a series of observations indicating that one's language and socioeconomic class are deeply intertwined. This realization actually provides a stepping stone to design systems able to infer the former from the latter. In this chapter, we hence demonstrate how semantic and syntactic features can be turned into training data for an ensemble classifier of socioeconomic status.

The quantification and inference of SES of individuals is a long-standing question in the social sciences. It is a rather difficult problem as it may depend on a combination of individual characteristics and environmental variables [146]. Some of these features can be easier assessed like income, gender, or age whereas others, relying to some degree on self-definition and sometimes entangled with privacy issues, are harder to assign like ethnicity, occupation, education level or home location. Furthermore, as we previously saw, individual SES correlates with other individual or network attributes, as users tend to build social links with others of similar SES, a phenomenon known as status homophily[157], arguably driving the observed stratification of society [135]. At the same time, shared social environment, similar education level, and social influence have been shown to jointly lead socioeconomic groups to exhibit stereotypical behavioral patterns, such as shared political opinion [27] or similar linguistic patterns [2]. Although these features are entangled and causal relation between them is far from understood, they appear as correlations in the data.

As the previous chapter proved, datasets recording multiple characteristics of human behaviour are more and more available due to recent developments in data collection technologies and increasingly popular online platforms and personal digital devices. The automatic tracking of online activities (commonly associated with profile data and meta-information), the precise recording of daily interaction dynamics and mobility patterns collected through mobile personal devices, together with the detailed and expert annotated census data all provide new grounds for the inference of individual features or behavioral patterns [132]. The exploitation of these data sources has already been proven to be fruitful as cutting edge recommendation systems, advanced methods for health record analysis, or successful prediction tools for social behaviour heavily rely on them [117]. Nevertheless, despite the available data, some inference tasks, like individual SES prediction, remain an open challenge.

The precise inference of SES would contribute to overcome multiple scientific challenges and could potentially have multiple commercial applications [136]. Further, robust SES inference would provide unique opportunities to gain deeper insights into socioeconomic inequalities [177], social stratification [135], and into the mechanisms driving network evolution, such as status homophily or social segregation.

In this chapter, we take a horizontal approach to this problem and explore various ways to infer the SES of a large sample of social media users. We propose different data collection and combination strategies using open, crawlable, or expert annotated socioeconomic data for the prediction task. Specifically, we use an extensive Twitter dataset of 1.3M users located in France, all associated with their tweets and profile information; 32,053 of them having inferred home locations. This collection actually results from the subsampling and extension from the one previously used Individual SES is estimated by relying on three separate datasets, namely socioeconomic census data; crawled profession information; and expert annotated Google Street View images of users' home locations. Each of these datasets is then used as ground-truth to infer the SES of Twitter users from profile and semantic features similar to [181]. We explore and assess how the SES of social media users can be obtained and how much the inference problem depends on annotation and the user's individual and linguistic attributes. In addition, to demonstrate the power of our location inference method, we group users into nine distinct socioeconomic classes to identify correlations between the predictability of mobility [207] of geolocated users and their socioeconomic status. We observe that as a user's SES increases, so to does his/her radius of gyration which in turn lowers the upper bound of predictability of his/her whereabouts.

We provide in Section 3.2 an overview of the related literature to contextualize the novelty of our work. In Section 3.3 we provide a detailed description of the data

collection and combination methods including analysis of Twitter, census, mobility, occupation and home location data. In Section 3.4 we introduce the features extracted to solve the SES inference problem, with results summarized in Section 3.5.

3.2 Related Work

There is a growing effort in the field of computational social science to combine online behavioral data with census records, and expert annotated information to infer social attributes of users of online services. The predicted attributes range from easily assessable individual characteristics such as age [35], or occupation [181, 182, 98] to more complex psychological and sociological traits like political affiliation [225], personality [199], or SES [152, 181].

Predictive features proposed to infer the desired attributes are also numerous. In case of Twitter, user information can be publicly queried within the limits of the public API [224]. User characteristics collected in this way, such as profile features, tweeting behavior, social network and linguistic content have been used for prediction, while other inference methods relying on external data sources such as website traffic data [44] or census data [149, 54] have also proven effective. Nonetheless, only recent works involve user semantics in a broader context related to social networks, spatiotemporal information, and personal attributes [182, 181, 6].

In this framework, aggregated studies of user spatial data have been particularly useful in fueling the analysis of human mobility patterns. Early work on mobile communication datasets [207, 70] showed that individuals tend to return to a few highly frequented locations leading to a high predictability in human travelling patterns. Analogous behaviour was later exposed in Twitter [109], enabling the use of this social media platform as a proxy for tracking and predicting human movement.

Akin to this strand of research, user features collected from online platforms have also been used in the inference of individual demographic traits. The tradition of relating SES of individuals to their language dates back to the early stages of sociolinguistics where it was first shown that social status reflected through a person's occupation is a determinant factor in the way language is used [17]. This line of research was recently revisited by Lampos *et al.* to study the SES inference problem on Twitter. In a series of works [182, 181, 6], the authors applied Gaussian Processes to predict user income, occupation and socioeconomic class based on demographic, psycho-linguistic features and a standardized job classification taxonomy, which mapped Twitter users to their professional occupations. The high predictive performance has proven this concept with $r = 0.633$ for income prediction, and a precision of 55% for 9-ways SOC

classification, and 82% for binary SES classification. Nevertheless, the models developed by the authors are learned by relying on datasets, which were manually labeled through an annotation process crowdsourced through Amazon Mechanical Turk at a high monetary cost. Although the labeled data has been released and provides the base for new extensions [35], it has two potential shortfalls that need to be acknowledged. First, the method requires access to a detailed job taxonomy, in this case specific to England, which hinders potential extensions of this line of work to other languages and countries. Furthermore, the language to income pipeline seems to show some dependency on the sample of users that actively chose to disclose their profession in their Twitter profile. Features obtained on this set might not be easily recovered from a wider sample of Twitter users. This limits the generalization of these results without assuming a costly acquisition of a new dataset.

3.3 Data collection and combination

Our first motivation in this study was to overcome earlier limitations by exploring alternative data collection and combination methods. We study here three ways to estimate the SES of Twitter users by using (a) open census data, (b) crawled and manually annotated data on professional skills and occupation, and (c) expert annotated data on home location Street View images. We provide here a collection of procedures that enable interested researchers to introduce predictive performance and scalability considerations when interested in developing language to SES inference pipelines. In the following we present in detail all of our data collection and combination methods.

3.3.1 Twitter corpus

In this work, we rely significantly upon the datasets previously collected to build several different corpora:

Geolocated users

To find users with a representative home location we followed the method published in [41, 97]. As a bottom line, we concentrated on $127,614$ users who posted at least five geolocated tweets with valid GPS coordinates, with at least three of them within a valid census cell (for definition see later), and over a longer period than seven days. Applying these filters we obtained 1,000,064 locations from geolocated tweets. By focusing on the geolocated users, we kept those with limited mobility, i.e., with median distance between locations not greater than 30 km, with tweets posted at places and times, which did not require travel faster than 130 km/h (maximum speed allowed within France), and with no more than three tweets within a two seconds

window. We further filtered out tweets with coordinates corresponding to locations referring to places (such as “Paris" or “France"). Thus, we removed locations that didn't exactly correspond to GPS-tagged tweets and also users, which were most likely bots. Despite the inherent similarities in deriving home location with respect to the previous study, the current methodology imposes a more stringent filtering on raw geolocation data to derive a more finegrained and accurate home location and socioeconomic status assignment.

Home location was estimated by the most frequent location for a user among all coordinates she visited. This way we obtained $32,053$ users, each associated with a unique home location. Finally, we collected the latest $3,200$ tweets from the timeline of all of geolocated users using the Twitter public API [224]. Note, that by applying these consecutive filters we obtained a more representative population as the Gini index, indicating overall socioeconomic inequalities, was 37.3% before filtering become 36.4% due to the filtering methods, which is closer to the value reported by the World Bank (33.7%) [10]. To verify our results, we computed the average weekday and weekend distance from each recorded location of a user to her inferred home location defined either as her most frequent location overall or among locations posted outside of work-hours from 9AM to 6PM (see Fig. 3-1a and b)). This circadian pattern displays great similarity to earlier results [97] with two maxima, roughly corresponding to times at the workplace, and a local minimum at 1PM due to people having lunch at home for locations posted in weekdays. Moreover, when comparing the weekday and weekend patterns of behaviour, we notice that the average distance to the inferred home location seemed to be smaller when considering only geolocations posted during the weekend, most likely due to the absence of the home-to-work commuting during the weekend. We found that this circadian pattern was more consistent with earlier results [97] when we considered all geolocated tweets (“All" in Fig. 3-1a) rather than only tweets including “home-related" expressions (“Night" in Fig. 3-1a). To further verify the inferred home locations, for a subset of 29,389 users we looked for regular expressions in their tweets that were indicative of being at home [97], such as “chez moi", “bruit", “dormir" or “nuit". In Fig. 3-1c we show the temporal distribution of the rate of the word “dormir" at the inferred home locations. This distribution appears with a peak around 10PM, which is very different from the overall distribution of geolocated tweets throughout the day considering any location (see Fig. 3-1b).

Linguistic data

To obtain meaningful linguistic data we pre-processed the incoming tweet streams in several ways. As our central question here deals with language semantics of individuals, re-tweets do not bring any additional information to our study, thus we removed them by default. We also removed any expressions considered to be seman-

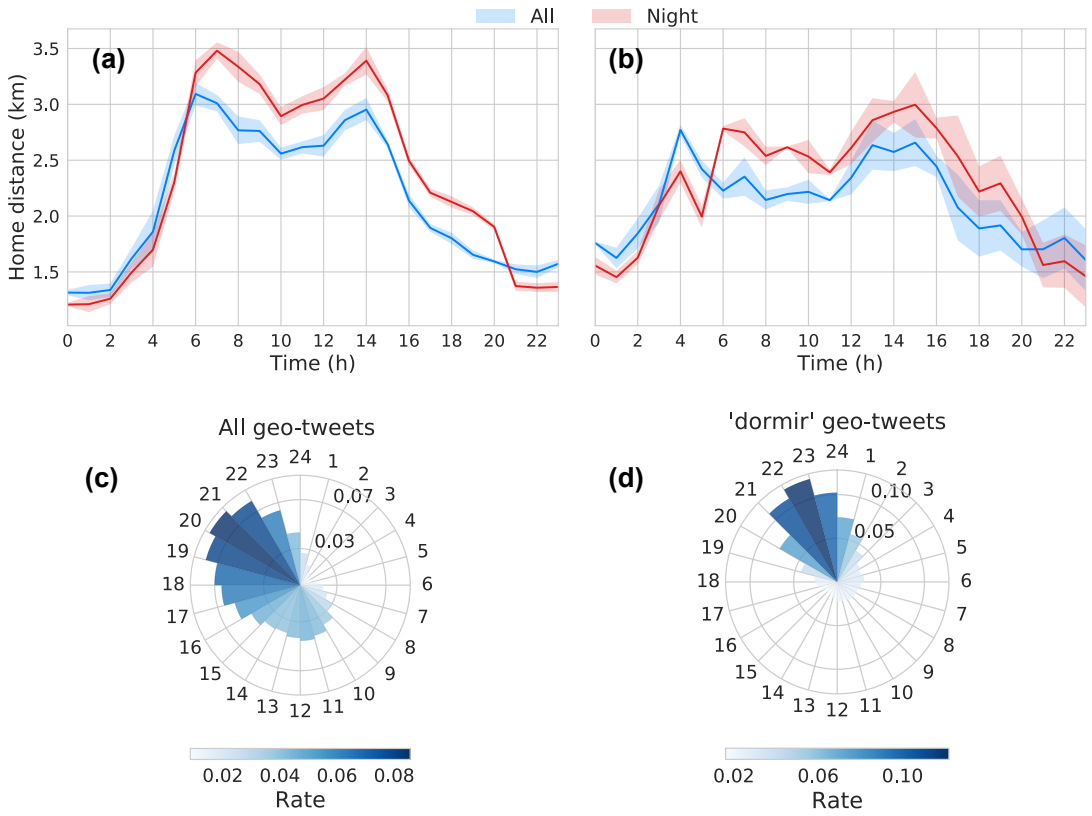


Figure 3-1: Average distance from home of active users per hour of the day depending on whether the geolocated tweet was posted during in a weekday (a) or the weekend (b) and the most frequent location was chosen among all (blue) or only the night ones (red). (c) Hourly rate of all geolocated tweets and (d) geolocated tweets mentioning 'dormir' averaged over all weekdays.

tically meaningless like URLs, emoticons, mentions of other users (denoted by the `@` symbol) and hashtags (denoted by the `#` symbol) to simplify later post-processing. In addition, as a last step of textual pre-processing, we downcased and stripped the punctuation from the text of every tweet. The reader might be wondering why the current algorithm doesn't build upon exclusively on the previous correlations to derive SES inference. The reasons for these are two-fold: As it turns out, we observed that while such correlations are meaningful at the population level the former study was conducted, they hold little predictive power of SES when studying users at the individual level. Furthermore, the here proposed way of deriving semantical information enables the extraction of these and other richer linguistic features that, as we'll be able to see, later permitted the completion of our inference task.

3.3.2 Census data

Our first method to associate SES to geolocated users builds on an open census income dataset at intra-urban level for France [105]. For this study, we moved away from the 200m x 200m dataset and relied upon a coarser grained but more discriminative sociodemographic data. In doing so, we hoped to avoid the biases associated with the winsorization of the income distribution. This dataset collects detailed socioeconomic information of individuals at the census block level (called IRIS), which are defined as territorial cells with varying size but corresponding to blocks of around $2,000$ inhabitants, as shown in Fig. 3-2 for greater Paris. For each cell, the data records the deciles of the income distribution of inhabitants. Note that the IRIS data does not provide full coverage of the French territory, as some cells were not reported to avoid identification of individuals (in accordance with current privacy laws), or to avoid territorial cells of excessive area. Nevertheless, this limitation did not hinder our results significantly as we only considered users who posted at least three times from valid IRIS cells, as explained in Section 3.3.1. To associate a single income value to

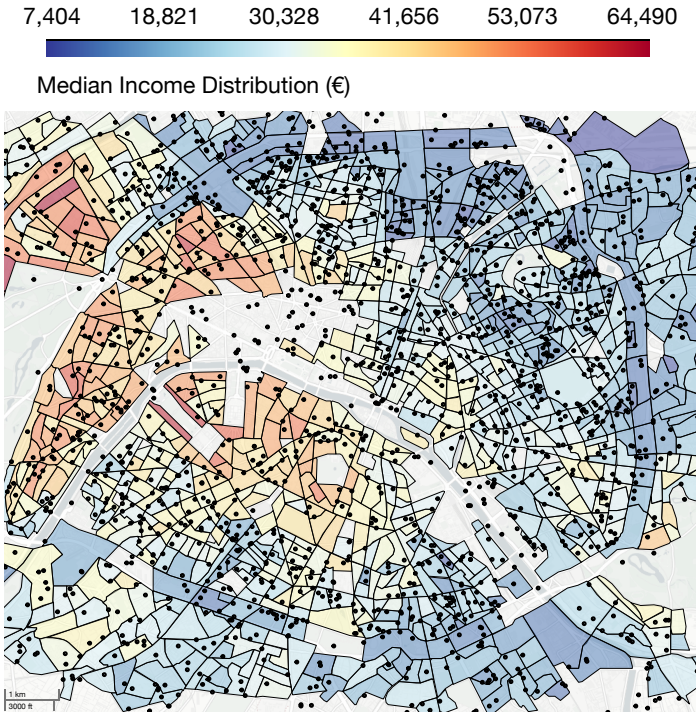


Figure 3-2: IRIS area cells in central Paris colored according to the median income of inhabitants, with inferred home locations of $2,000$ Twitter users.

each user, we identified the cell of their estimated home locations and assigned them with the median of the corresponding income distribution. Thus we obtained an average socioeconomic indicator for each user, which was distributed heterogeneously in accordance with Pareto's law [170]. This is demonstrated in Fig. 3-6a, where the $C(f)$ cumulative income distributions as the function of population fraction f ap-

pears as a Lorentz-curve with area under the diagonal proportional to socioeconomic inequalities. As an example, Fig. 3-2 depicts the spatial distribution of 2,000 users with inferred home locations in IRIS cells located in central Paris and colored as the median income.

3.3.3 Mobility Analysis

In order to further assess the validity of the assignment of location to socioeconomic status, we studied in this section the mobility traces generated by the set of geolocated users. Specifically, mirroring previous work [207, 109], we focused on the predictability of individual trajectories by analyzing the visitation patterns of the top $n = 10$ locations. To do so, given a user having visited at least $k = 5$ different census blocks, we computed $\pi_k(i)$ the fraction of time a user spends in the top i most visited location as:

$$\pi_k(i) = \frac{N_i}{\sum_{j=1}^{n} N_j}, \quad i = 1, ..., n$$

with N_i the number of times the user appeared in the i-th location. Note that in the above definition, $\forall i \leq k, \quad N_i \in \mathbb{N}^*$. This metric was shown by Song et al. [207] to be an upper bound to an individual's predictability. At the same time, taking geolocated users' median income inferred from census cells we associated them into one of nine socioeconomic classes (1-poorest, 9-richest) following the social stratum model introduced in [135]. This procedure sorts users by their income, takes the CDF of income (shown in Fig. 3-6a) and divides users into groups such that each group represents the same total sum of income. This partitioning, due to the Lorentz-curve shape of the CDF, provides socioeconomic classes with decreasing size with increasing income, as expected from theory and other real observations [194, 135].

As previously pointed out [207, 109] we noticed, that for all socioeconomic classes, the first and second most visited location concentrate between 60 and 74% of all geolocations , suggesting the high likelihood of an individual tweeting preferentially from his/her home or office [109]. Furthermore, by aggregating users per socioeconomic class, an interesting trend was exposed: the higher the socioeconomic class in question, the lower the (upper bound of) predictability of the considered users was for any of the top n locations. Indeed, when studying how $\pi_5(1)$ is related to a user's socioeconomic status, we see that the higher one's socioeconomic class the less frequently he is seen at the most visited location (see Fig. 3-3a). The robustness of this trend was further assessed by repeating our analysis for different values of k, always recovering this decreasing trend (see Fig. 3-3b). The underlying explanation to this pattern may actually lie in the correlation between the socioeconomic status of people and the diversity of locations they visit. For instance, greater diversity may lead to a lower predictability of movement, which in turn might cause the observed

trends. To control for this, we computed the average radius of gyration r_g, describing the typical range of a user's trajectory per socioeconomic class, defined as:

$$r_g = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(\overrightarrow{r_i} - \overrightarrow{r_{cm}})^2}.$$

Here $\overrightarrow{r_i}$ represents the position at time i, $\overrightarrow{r_{cm}} = \frac{1}{L}\sum_{i=1}^{L}\overrightarrow{r_i}$ is the center of mass of the trajectory and L is the total number of recorded points for the userâĂŹs location. As we see in Fig. 3-3c, r_g seems to increase on average as higher socioeconomic status is considered. Hence, high SES users tend to have more diverse mobility patterns than low SES ones, which in turn leads to lower predictability of their whereabouts. These results may relate to previous work [229, 135], which explain this trend by means of the positive payoff between commuting farther for better jobs, while keeping better housing conditions. As a consequence, the inferred home location for high SES users might be less precise due to their more dispersed mobility patterns and the lower predictability of their whereabouts. This is one reason why we define our SES inference later as a two-ways inference problem, dividing users into a "rich" and a "poor" class.

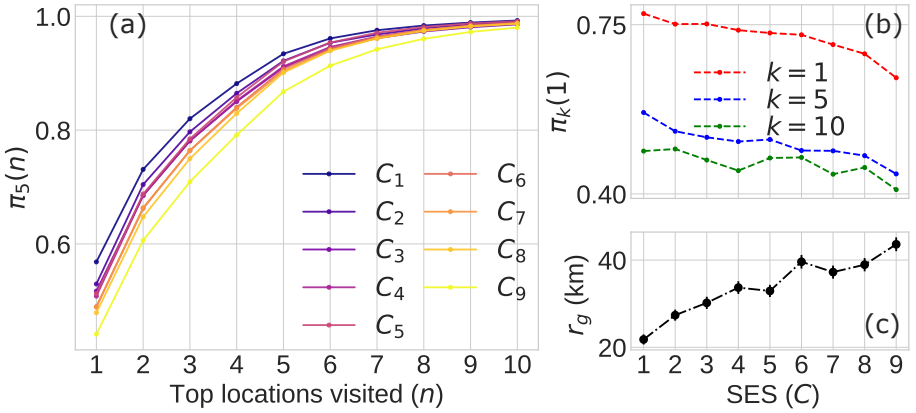


Figure 3-3: (a) The fraction of time a user spends in his/her top 10 most visited locations per socioeconomic class. (b) $\pi_k(1)$, fraction of time spent at the most visited location per socioeconomic class for users with at least k different locations. (c) Average radius of gyration per socioeconomic class (error-bars are computed as $\sigma(r_g)/\sqrt{N}$, with N number of samples)

3.3.4 Occupation data

Earlier studies [181, 182] demonstrated that annotated occupation information can be effectively used to derive precise income for individuals and to infer therefore their SES. However, these methods required a somewhat selective set of Twitter users

as well as an expensive annotation process by hiring premium annotators e.g. from Amazon Mechanical Turk. Our goal here was to obtain the occupations for a general set of Twitter users without the involvement of annotators, but by collecting data from parallel online services.

As a second method to estimate SES, we took a sample of Twitter users who mentioned their LinkedIn [148] profile url in their tweets or Twitter profile. Using these pointers we collected professional profile descriptions from LinkedIn by relying on an automatic crawler mainly used in Search Engine Optimization (SEO) tasks [1]. We obtained $4,140$ Twitter/LinkedIn users all associated with their job title, professional skills and profile description. Apart from the advantage of working with structured data, professional information extracted from LinkedIn is significantly more reliable than Twitter's due to the high degree of social scrutiny to which each profile is exposed [153].

To associate income to Twitter users with LinkedIn profiles, we matched them with a given salary based on their reported profession and an occupational salary classification table provided by INSEE [103]. Due to the ambiguous naming of jobs and to acknowledge permanent/non-permanent, senior/junior contract types we followed three strategies for the matching. In 40% of the cases we directly associated the reported job titles to regular expressions of an occupation. In 50% of the cases we used string sequencing methods [59] to associate reported and official names of occupations with at least 90% match. For the remaining 10% of users we directly inspected profiles. The distribution of estimated salaries reflects the expected income heterogeneities as shown in Fig. 3-6a. Users were eventually assigned to one of two SES classes based on whether their salary was higher or lower than the average value of the income distribution. Also note, that LinkedIn users may not be representative of the whole population. We discuss this and other types of potential biases in Section 3.6.

3.3.5 Expert annotated home location data

Finally, motivated by recent remote sensing techniques, we sought to estimate SES via the analysis of the urban environment around the inferred home locations. Similar methodology has been lately reported by the remote sensing community [63] to predict socio-demographic features of a given neighborhood by analyzing Google Street View images to detect different car models, or to predict poverty rates across urban areas in Africa from satellite imagery [108]. Driven by this line of work, we estimated the SES of geolocated Twitter users as follows:

Pre-selection of home locations

Using geolocated users identified in Section 3.3.1, we further filtered them to obtain a smaller set of users with more precise inferred home locations. We screened all of their geotagged tweets and looked for regular expressions determining whether or not a tweet was sent from home [97]. As explained in Section 3.3.1, we exploited that "home-suspected" expressions appeared with a particular temporal distribution (see Fig. 3-1c) since these expressions were used during the night when users are at home. This selection yielded 28,397 users mentioning "home-suspected" expressions regularly at their inferred home locations.

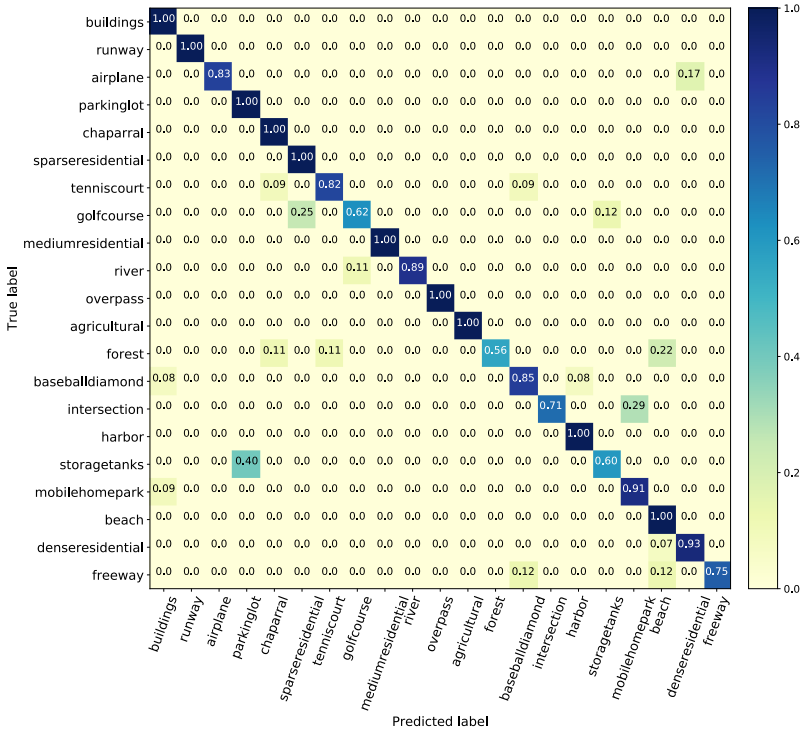


Figure 3-4: Confusion matrix in the test set obtained with a ResNet50 trained on the UCMerced for land use inference

Identification of urban/residential areas

In order to filter out inferred home locations not in urban/residential areas, we downloaded via Google Maps Static API [71] a satellite view in a $100m$ radius around each coordinate (for a sample see Fig. 3-5a). To discriminate between residential and non-residential areas, we built on land use classifier [33] using aerial imagery from the UC Merced dataset [230]. This dataset contains 2100 256×256 $1m/px$ aerial RGB images over 21 classes of different land use (for a pair of sample images see Fig. 3-5b). To classify land use, a CaffeNet architecture was trained, which reached an accuracy

over 95%. Here, we instantiated a *ResNet50* network using *keras* [38] pre-trained on ImageNet [48], where all layers except the last five were frozen. The network was then trained with 10-fold cross validation achieving a 93% accuracy after the first 100 epochs (cf. Figure 3-4). We used this model to classify images of the estimated home location satellite views (cf. Figure 3-5a) and kept those which were identified as residential areas (see Fig. 3-5b, showing the activation of the two first hidden layers of the trained model). This way 5,396 inferred home locations were discarded.

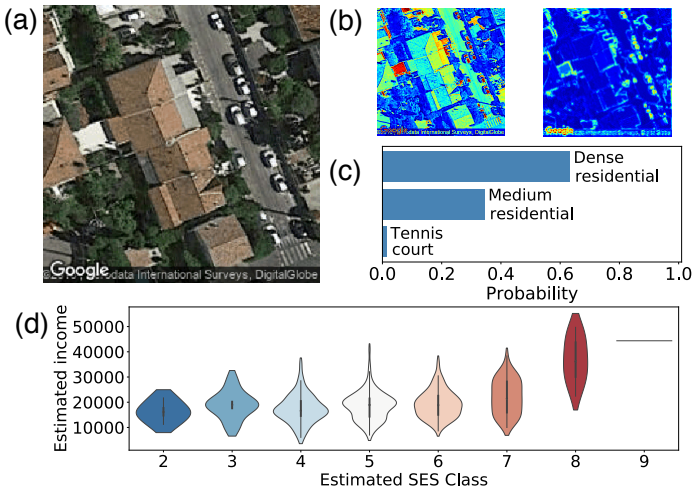


Figure 3-5: Top: ResNet50 Output: (a): Original satellite view; (b): First two hidden layers activation; (c): Final top-3 most frequent predicted area types; (d) Architect SES score agreement with census median income for the sampled home locations. It is shown as violin plots of income distributions for users annotated in different classes (shown on x-axis and by color).

Home location data with expert annotated SES

Next we aimed to estimate SES from architectural/urban features associated to the home locations. Given that our goal here was to lean on socioeconomic labels that weren't estimated by the census, we forfeit the use of deep learning models to infer SES by relying upon human annotation. Thus, for each home location we collected two additional satellite views at different resolutions as well as six Street View images, each with a horizontal view of approximately 90°. We randomly selected a sample of 1,000 locations and involved architects to assign a SES score (from 1 to 9) to a sample set of selected locations based on the satellite and Street View around it (both samples had 333 overlapping locations). For validation, we took users from each annotated SES class and computed the distribution of their incomes inferred from the IRIS census data (see Section 3.3.2). Violin plots in Fig. 3-5d show that in expert annotated data, as expected, the inferred income values were positively correlated with the annotated SES classes. Labels were then categorized into two socioeconomic

classes for comparison purposes. All in all, both annotators assigned the same label to the overlapping locations in 81.7% of samples.

	Census	**Occupation**	**Expert**
Size	32,053	4,140	1,000
Low SES	0.54	0.46	0.58
High SES	0.46	0.54	0.42

Table 3.1: Number of users and estimated fractions of low and high SES in each dataset

To solve the SES inference problem we used the above described three datasets (for a summary see Table 3.1). We defined the inference task as a two-way classification problem by dividing the user set of each dataset into two groups. For the census and occupation datasets the lower and higher SES classes were separated by the average income computed from the whole distribution, while in the case of the expert annotated data we assigned people from the lowest five SES labels to the lower SES class in the two-way task. The relative fractions of people assigned to the two classes are depicted in Fig. 3-6b for each dataset and summarized in Table 3.1.

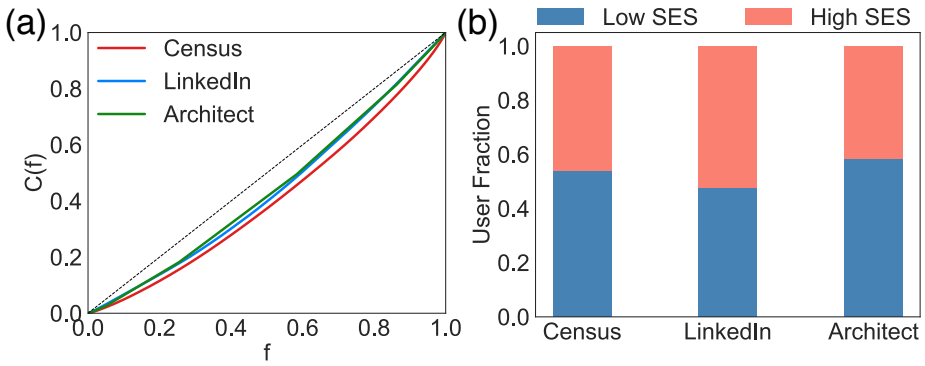


Figure 3-6: Cumulative distributions of income as a function of sorted fraction f of individuals. Dashed line corresponds to the perfectly balanced distribution. Distributions appear similar in spite of dealing with heterogeneous samples.

3.4 Feature selection

Using the user profile information and tweets collected from every account's timeline, we built a feature set for each user, similar to Lampos *et al.* [181]. We categorized features into two sets, one containing shallow features directly observable from the data, while the other was obtained via a pipeline of data processing methods to capture semantic user features.

3.4.1 User Level Features

The user level features are based on the general user information or aggregated statistics about the tweets [181]. We therefore include general ordinal values such as the number and rate of retweets, mentions, and coarse-grained information about the social network of users (number of friends, followers, and ratio of friends to followers). Finally we vectorized each user's profile description and tweets and selected the top 450 and 560 1-grams and 2-grams, respectively, observed through their accounts (where the rank of a given 1-gram was estimated via *tf-idf* [138]).

3.4.2 Linguistic features

To represent textual information, in addition to word count data, we used topic models to encode coarse-grained information on the content of the tweets of a user, similar to [181]. This enabled us to easily interpret the relation between semantic and socioeconomic features. Specifically, we started by training a *word2vec* model [160] on the whole set of tweets (obtained in the 2014-2015 time-frame) by using the skip-gram model and negative sampling with parameters similar to [35]. To scale up the analysis, the number of dimensions for the embedding was kept at 50. This embedded words in the initial dataset in a $\mathbb{R}^{50}$ vector space. Eventually we extracted conversation topics by running a spectral clustering algorithm on the word-to-word similarity matrix $M \in \mathbb{R}^{V\times V}$ with V vocabulary size and elements defined as the $M_{ij} = \frac{\langle u_i, u_j\rangle}{||u_i||||u_j||}$ cosine similarity between word vectors. Here $u_i \in \mathbb{R}^{50}$ is a vector of a word $i \in V$ in the embedding, $\langle\cdot\rangle$ is the dot product of vectors, and $||\cdot||$ is the L^2 norm of a vector. This definition allows for negative entries in the matrix to cluster, which were set to null in our case. This is consistent with the goal of the clustering procedure as negative similarities shouldn't encode dissimilarity between pairs of words but orthogonality between the embeddings. This procedure was run for 50, 100 and 200 clusters and allowed the homogeneous distribution of words among clusters (hard clustering). The best results were obtained with 100 topics in the topic model. Finally, we manually labeled topics based on the words assigned to them, and computed the topic-to-topic correlation matrix shown in Fig. 3-7. There, after block diagonalization, we found clearly correlated groups of topics which could be associated to larger topical areas such as communication, advertisement or soccer.

As a result we could compute a representative topic distribution for each user, defined as a vector of normalized usage frequency of words from each topic. Also note that the topic distribution for a given user was automatically obtained as it depends only on the set of tweets and the learned topic clusters without further parametrization.

To demonstrate how discriminative the identified topics were in terms of the SES of users we associated to each user the 9th decile value of the income distribution cor-

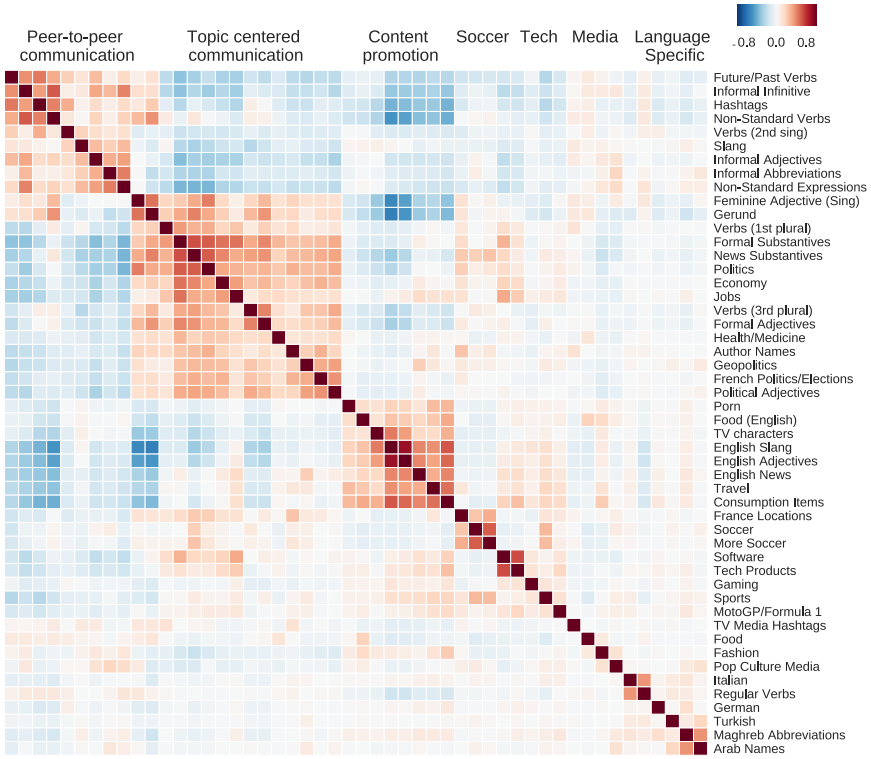


Figure 3-7: Clustered topic-to-topic correlation matrix: Topics are generated via the spectral clustering of the word2vec word co-similarity matrix. Row labels are the name of topics while column labels are their categories. Blue cells (resp. red) assign negative (resp. positive) Pearson's correlation coefficients.

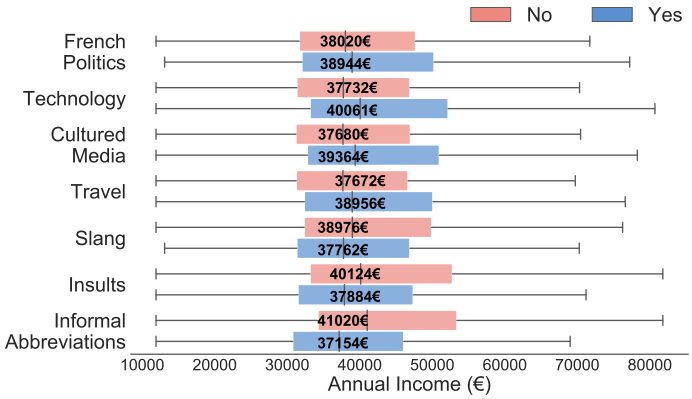


Figure 3-8: Average income for users who tweeted about a given topic (blue) vs. those who didn't (red). Label of the considered topic is on the left.

responding to the census block of their home location and computed for each labelled topic the average income of users depending on whether or not they mentioned the given topic. Results in Fig. 3-8 demonstrate that topics related to politics, technology or culture are more discussed by people with higher income, while other topics associated to slang, insults or informal abbreviations are more used by people of lower

income. These observable differences between the average income of people, who use (or not) words from discriminative topics, demonstrates well the potential of word topic clustering used as features for the inference of SES. All in all, each user in our dataset was assigned with a 1117 feature vector encoding the lexical and semantic profile she displayed on Twitter. We did not apply any further feature selection as the distribution of importance of features appeared rather smooth (not shown here). It did not provided evident ways to identify a clear set of particularly determinant features, but rather indicated that the combination of them were important.

3.5 Results

In order to assess the degree to which linguistic features can be used for discriminating users by their socioeconomic class, we trained with these feature sets different learning algorithms. Namely, we used the XGBoost algorithm [36], an implementation of the gradient-boosted decision trees for this task. Training a decision tree learning algorithm involves the generation of a series of rules, split points or nodes ordered in a tree-like structure enabling the prediction of a target output value based on the values of the input features. More specifically, XGBoost, as an ensemble technique, is trained by sequentially adding a high number of individually weak but complementary classifiers to produce a robust estimator: each new model is built to be maximally correlated with the negative gradient of the loss function associated with the model assembly [220]. To evaluate the performance of this method we benchmarked it against more standard ensemble learning algorithms such as AdaBoost, Logistic Regression, SVM and Random Forest. For each socioeconomic dataset, we

	Census	**Occupation**	**Expert**
AdaBoost	0.549 ± 0.009	0.628 ± 0.022	0.575 ± 0.013
Logistic Reg.	0.658 ± 0.013	0.778 ± 0.058	0.571 ± 0.033
SVM	0.657 ± 0.016	0.788 ± 0.041	0.600 ± 0.012
Random Forest	0.677 ± 0.011	0.783 ± 0.017	0.593 ± 0.049
XGBoost	$\mathbf{0.700 \pm 0.011}$	$\mathbf{0.798 \pm 0.015}$	$\mathbf{0.605 \pm 0.029}$

Table 3.2: Classification performance (5-CV): AUC scores (mean ± STD) of five different classifiers on each dataset)

trained our models by using 75% of the available data for training and the remaining 25% for testing. During the training phase, the training data undergoes a k-fold inner cross-validation, with $k = 5$, where all splits are computed in a stratified manner to get the same ratio of lower to higher SES users. The four first blocks were used for inner training and the remainder for inner testing. This was repeated ten times for each model so that in the end, each model's performance on the validation set was

averaged over 50 samples. For each model, the parameters were fine-tuned by training 500 different models over the aforementioned splits. The selected one was that which gave the best performance on average, which was then applied to the held-out test set. This is then repeated through a 5-fold outer cross-validation.

In terms of prediction score, we followed a standard procedure in the literature [185] and evaluated the learned models by considering the area under the receiver operating characteristic curve (AUC). This metric can be thought as the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one [220]. This procedure was applied to each of our datasets. The obtained results are shown in Fig. 3-9 and in Table 3.3. As a result, we first observed that

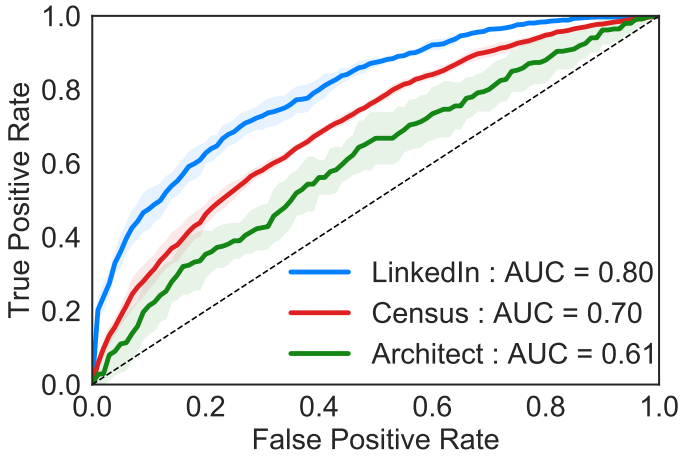


Figure 3-9: ROC curves for 2-way SES prediction using tuned XGBoost in each of the 3 SES datasets. AUC values are reported in the legend. The dashed line corresponds to the line of no discrimination. Solid lines assign average values over all folds while shaded regions represent standard deviation.

XGBoost consistently provided top prediction scores when compared to AdaBoost and Random Forest (all performance scores are summarised in Table 3.2). We hence used it for our predictions in the remainder of this study. We found that the LinkedIn data was the best, with $AUC = 0.80$, to train a model to predict SES of people based on their semantic features. It provided a 10% increase in performance as compared to the census based inference with $AUC = 0.70$, and 19% relative to expert annotated data with $AUC = 0.61$. Thus we can conclude that there seem to be a trade-off between scalability and prediction quality, as while the occupation dataset provided the best results, it seems unlikely to be subject to any upscaling due to the high cost of obtaining a clean dataset. Relying on location to estimate SES seems to be more likely to benefit from such an approach, though at the cost of an increased number of mislabelled users in the dataset. Moreover, the annotator's estimation of SES using Street View at each home location seems to be hindered by the large variability of

urban features. Note that even though inter-agreement is 76%, the Cohen's kappa score for annotator inter-agreement is low at 0.169. Furthermore, we remark that the expert annotated pipeline was also subject to noise affecting the home location estimations, which potentially contributed to the lowest predictive performance. We

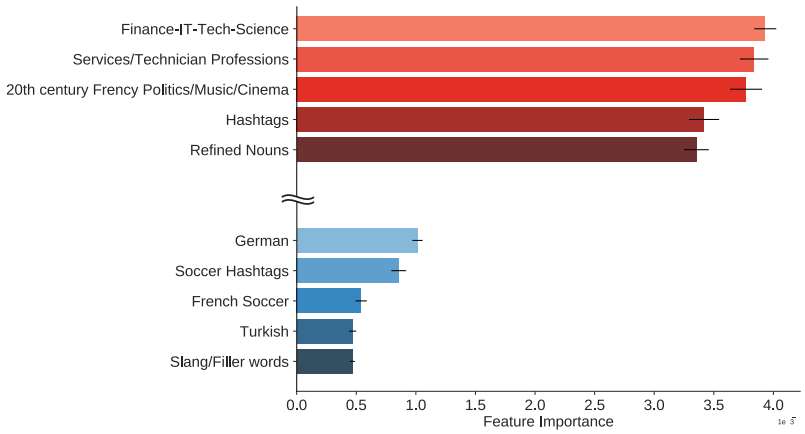


Figure 3-10: Top (red) and bottom (blue) five topics ranked in terms of their predictive performance in the XGBoost model trained on LinkedIn. Error bars indicate s.d. across the 10 cross-validation samples.

also report the top and bottom five topics ranked by their importance when the XGBoost model was trained on the best performing proxy, i.e., on occupation (see Figure 3-10). Perhaps unsurprisingly, topics related to professional occupations are the ones recognized by the model as most important. Nevertheless, syntax remains an important feature too. Furthermore, topics associated to particular communities (German/Turkish) or general interest (Soccer) seem to be less useful in terms of SES discrimination. This could in turn be explained by the sparsity of individuals using them or inversely, by the breadth of users discussing them.

Dataset	SES Class	Performance on test set		
		Precision	Recall	F1-score
Census	Low	0.652	0.596	0.624
	High	0.628	0.682	0.652
LinkedIn	Low	0.700	0.733	0.717
	High	0.735	0.702	0.720
Architect	Low	0.622	0.598	0.607
	High	0.550	0.573	0.556

Table 3.3: Detailed average performance (5-CV) on test data for the binary SES inference problem for each of the 3 datasets

3.6 Limitations

In this work we combined multiple datasets collected from various sources. Each of them came with some bias due to the data collection and post-treatment methods or the incomplete set of users. These biases may limit the success of our inference, thus their identification is important for the interpretation and future developments of our framework.

• *Location data*: Although we designed very strict conditions for the precise inference of home locations of geolocated users, this process may have some uncertainty due to outlier behaviour. Further bias may be induced by the relatively long time passed between the posting of the location data and of the tweets collection of users.

• *Census data*: As we already mentioned the census data does not cover the entire French territory as it reports only cells with close to $2,000$ inhabitants. This may introduce biases in two ways: by limiting the number of people in our sample living in rural areas, and by associating income with large variation to each cell. While the former limit had marginal effects on our predictions, as Twitter users mostly live in urban areas, we addressed the latter effect by associating the median income to users located in a given cell.

• *Occupation data*: LinkedIn as a professional online social network is predominantly used by people from IT, business, management, marketing or other expert areas, typically associated with higher education levels and higher salaries. Moreover, we could observe only users who shared their professional profiles on Twitter, which may further biased our training set. In terms of occupational-salary classification, the data in [103] was collected in 2010 thus may not contain more recent professions. These biases may induce limits in the representativeness of our training data and thus in the predictions' precision. However, results based on this method of SES annotation performed best in our measurements, indicating that professions are among the most predictive features of SES, as has been reported in [181].

• *Annotated home locations*: The remote sensing annotation was done by experts and their evaluation was based on visual inspection and biased by some unavoidable subjectivity. Although their annotations were cross-referenced and found to be consistent, they still contained biases, like over-representative middle classes, which somewhat undermined the prediction task based on this dataset.

• *Different sets of users*: Our methodologies rely on non-entirely overlapping user sets when they turn to SES inference using occupational data, census data, or remotely sensed values as proxy for individual socioeconomic status. The obtained results are undoubtedly linked to the set of individuals used in each dataset, which may affect the discriminative analysis on the advantages that each proxy provides for the inference task. On the other hand, due to the same collection filters and pre-processing conditions, users in these subsets may be considered similar enough to be able to compare the performance provided by the different methods.

Despite these shortcomings, using all three datasets, we were able to infer SES with performances close to earlier reported results, which were based on more thoroughly annotated datasets. Our results, and our approach of using open, crawlable, or remotely sensed data highlights the potential of the proposed methodologies.

3.7 Conclusions

This chapter introduced a novel methodology for the inference of the SES of Twitter users. We built our models combining information obtained from numerous sources, including Twitter, census data, LinkedIn and Google Maps. We developed precise methods of home location inference from geolocation, novel annotation of remotely sensed images of living environments, and effective combination of datasets collected from multiple sources. In terms of novelty, we demonstrated that within the French Twitter space, the utilization of words in different topic categories, identified via advanced semantic analysis of tweets, can discriminate between people of different income; and that the mobility patterns and such the predictability of users' whereabouts is strongly dependent on SES of people. Furthermore, we showed that among the candidate socioeconomic proxies chosen, the best results were obtained using occupational data. More importantly, we presented a proof-of-concept that our methods are competitive in terms of SES inference when compared to other methods relying on domain specific information. Note that after training, our model requires as input only information that can be collected exclusively from Twitter, without relying on other data sources. This holds a large potential in terms of SES inference of larger sets of Twitter users, which in turn opens the door for studies to address population level correlations of SES with language, space, time, or social network. In doing so it opens the gate to the greater development and refinement of precise SES inference techniques from disparate data sources. In the chapters that will follow, we will elaborate on this theme and show that deep learning tools are fully capable of extracting meaningful socioeconomic information from satellite imagery. Moreover, we will build upon the so far shown correlations between individual social features and collective network structure, to examine whether a model able to jointly represent node features and network structure has an enhanced representation power.

Chapter 4

Joint embedding of features and network structure with graph convolutional networks

This work is currently in submission as a journal paper in Applied Network Science [137] with myself as co-author

4.1 Introduction

In this chapter we bring together the threads of the previous two chapters and simultaneously consider network and user information (both features and target, e.g respectively user's language and socioeconomic status) as different dimensions of the same latent space. More precisely, in the triangle T defined by a user's network, language and socioeconomic status, we have so far confirmed the existence of pairwise correlations between all the concerned pairs and have shown how, at least for one of them, it was possible to infer it from the other. Here, we work under the underlying idea that these correlations are indeed evidence of hidden social mechanisms driving the variations observed in Chapter 2, hence enabling inference tasks between all pairs in T. Rather than carrying out the empirical validation of this hypothesis, we take a step back and ask what is exactly gained by modelling all these elements under the same framework. We expand this idea in greater detail below.

Although it is relatively easy to obtain the proxy social network and various individual features for users of online social platforms, the combined characterisation of these types of information is still challenging our methodology. While current approaches have been able to approximate the observed marginal distributions of node and network features separately, their combined consideration was usually done via summary network statistics merged with otherwise independently built feature sets

of nodes. However, the entanglement between structural patterns and feature similarities appears to be fundamental to a deeper understanding of network formation and dynamics. The value of this joint information then calls for the development of statistical tools for the learning of combined representation of network and feature information and their dependencies.

The formation of ties and mesoscopic structures in online social networks is arguably determined by several competing factors. Considering only network information, neighbour similarities between people are thought to explain network communities, where triadic closure mechanisms [125, 118] induce ties between peers with larger fractions of common friends [74]. Meanwhile, random bridges [118] are built via focal closure mechanisms, optimising the structure for global connectedness and information dissemination. At the same time, people in the network can be characterised by various individual features such as their socio-demographic background [135, 2], linguistic characters [77, 2, 141], or the distributions of their topics of interests [34, 94], to mention just a few. Such features generate homophilic tie creation preferences [157, 119], which induce links with higher probability between similar individuals, whom in turn form feature communities of shared interest, age, gender, or socio-economic status, and so on [135, 204]. Though these mechanisms are not independent and lead to correlations between feature and network communities, it is difficult to define the causal relationship between the two: first, because simultaneously characterising similarities between multiple features and a complex network structure is not an easy task; second, because it is difficult to determine, which of the two types of information, features or structure, is driving network formation to a greater extent. Indeed, we do not know what fraction of similar people initially get connected through homophilic tie creation, versus the fraction that first get connected due to structural similarities before influencing each other to become more similar [126, 7].

Over the last decade popular methods have been developed to characterise structural and feature similarities and to identify these two notions of communities. The detection of network communities has been a major challenge in network science with various concepts proposed [19, 190, 175] to solve it as an unsupervised learning task [56, 57]. Commonly, these algorithms rely solely on network information, and their output is difficult to cross-examine without additional meta-data, which is usually disregarded in their description. On the other hand, methods grouping similar people into feature communities typically ignore network information, and exclusively rely on individual features to solve the problem as a data clustering challenge [110, 60]. Some semi-supervised learning tasks, such as link prediction, may take feature and structural information simultaneously into account, but only by enriching individual feature vectors with node-level network characteristics such as degree or local clustering [145, 150, 94]. Methods that would take higher order network corre-

lations and multivariate feature information into account at the same time are still to be defined. Their development would offer huge potential in understanding the relation between individuals' characters, their social relationships, the content they are engaged with, and the larger communities they belong to. This would not only provide us with deeper insight about social behaviour, it would give us predictive tools for the emergence of network structure, individual interests and behavioural patterns.

In this chapter we propose a contribution to solve this problem by developing a joint feature-network embedding built on multitask Graph Convolutional Networks [116, 28, 80, 231] and Variational Autoencoders (GCN-VAE) [112, 189, 115, 227, 233], which we call the Attributed Node to Vector method (AN2VEC). In our model, different dimensions of the generated embeddings can be dedicated to encode feature information, network structure, or shared feature-network information separately. Unlike previous embedding methods dealing with features [61, 222, 203, 21], this interaction model [3] allows us to explore the dependencies between the disentangled network and feature information by comparing the embedding reconstruction performance to a baseline case where no shared information is extracted. Using this method, we can identify an optimal reduced embedding, which indicates whether combined information coming from the structure and features is important, or whether their non-interacting combination is sufficient for reconstructing the featured network.

In practice, as this method solves a reconstruction problem, it may give important insights about the combination of feature- and structure-driven mechanisms which determine the formation of a given network. As an embedding, it is useful to identify people sharing similar individual and structural characters. And finally, by measuring the optimal overlap between feature– and network-associated dimensions, it can be used to verify network community detection methods to see how well they identify communities explained by feature similarities.

In what follows, after summarising the relevant literature, we introduce our method and demonstrate its performance on synthetic featured networks, for which we control the structural and feature communities as well as the correlations between the two. As a result, we will show that our embeddings, when relying on shared information, outperform the corresponding reference without shared information, and that this performance gap increases with the correlation between network and feature structure since the method can capture the increased joint information. Next, we extensively explore the behaviour of our model on link prediction and node classification on standard benchmark datasets, comparing it to well-known embedding methods.

4.2 Related Work

The advent of increasing computational power coupled with the continuous release and ubiquity of large graph-structured datasets has triggered a surge of research in the field of network embeddings. The main motivation behind this trend is to be able to convert a graph into a low-dimensional space where its structural information and properties are maximally preserved [29]. The aim is to extract unseen or hard to obtain properties of the network, either directly or by feeding the learned representations to a downstream inference pipeline.

4.2.1 Graph embedding survey: from matrix factorisation to deep learning

In early work, low-dimensional node embeddings were learned for graphs constructed from non-relational data by relying on matrix factorisation techniques. By assuming that the input data lies on a low dimensional manifold, such methods sought to reduce the dimensionality of the data while preserving its structure, and did so by factorising graph Laplacian eigenmaps [218] or node proximity matrices [31].

More recent work has attempted to develop embedding architectures that can use deep learning techniques to compute node representations. DeepWalk [176], for instance, computes node co-occurrence statistics by sampling the input graph via truncated random walks, and adopts a SkipGram neural language model to maximise the probability of observing the neighbourhood of a node given its embedding. By doing so the learned embedding space preserves second order proximity in the original graph. However, this technique and the ones that followed [76, 143] present generalisation caveats, as unobserved nodes during training cannot be meaningfully embedded in the representation space, and the embedding space itself cannot be generalised between graphs. Instead of relying on random walk-based sampling of graphs to feed deep learning architectures, other approaches have used the whole network as input to autoencoders in order to learn, at the bottleneck layer, an efficient representation able to recover proximity information [227, 32, 222]. However, the techniques developed herein remained limited due to the fact that successful deep learning models such as convolutional neural networks require an underlying euclidean structure in order to be applicable.

4.2.2 Geometric deep learning survey: defining convolutional layers on non-euclidean domains

This restriction has been gradually overcome by the development of graph convolutions or Graph Convolutional Networks (GCN). By relying on the definition of

convolutions in the spectral domain, Bruna et al. [28] defined spectral convolution layers based on the spectrum of the graph Laplacian. Several modifications and additions followed and were progressively added to ensure the feasibility of learning on large networks, as well as the spatial localisation of the learned filters [25, 231]. A key step is made by [47] with the use of Chebychev polynomials of the Laplacian, in order to avoid having to work in the spectral domain. These polynomials, of order up to r, generate localised filters that behave as a diffusion operator limited to r hops around each vertex. This construction is then further simplified by Kipf and Welling by assuming among others that $r \approx 2$ [116].

Recently, these approaches have been extended into more flexible and scalable frameworks. For instance, Hamilton et al. [80] extended the original GCN framework by enabling the inductive embedding of individual nodes, training a set of functions that learn to aggregate feature information from a node's local neighborhood. In doing so, every node defines a computational graph whose parameters are shared for all the graphs nodes.

More broadly, the combination of GCN with autoencoder architectures has proved fertile for creating new embedding methods. The introduction of probabilistic node embeddings, for instance, has appeared naturally from the application of variational autoencoders to graph data [189, 112, 115], and has since led to explorations of the uncertainty of embeddings [21, 233], of appropriate levels of disentanglement and overlap [154], and of better representation spaces for measuring pairwise embedding distances (see in particular recent applications of the Wasserstein distance between probabilistic embeddings [233, 163]). Such models consistently outperform earlier techniques on different benchmarks and have opened several interesting lines of research in fields ranging from drug design [52] to particle physics [114]. Most of the more recent approaches mentioned above can incorporate node features (either because they rely on them centrally, or as an add-on). However, with the exception of DANE [61], they mostly do so by assuming that node features are an additional source of information, which is congruent with the network structure (e.g. multi-task learning with shared weights [222], or fusing both information types together [203]). That assumption may not hold in many complex datasets, and it seems important to explore what type of embeddings can be constructed when we lift it, considering different levels of congruence between a network and the features of its nodes.

We therefore set out to make a change to the initial GCN-VAE in order to: (i) create embeddings that are explicitly trained to encode both node features and network structure; (ii) make it so that these embeddings can separate the information that is shared between network and features, from the (possibly non-congruent) information that is specific to either network or features; and (iii) be able to tune the

importance that is given to each type of information in the embeddings.

4.3 Methods

In this section we present the architecture of the neural network model we use to generate shared feature-structure node embeddings[1]. We take a featured network as input, with structure represented as an adjacency matrix and node features represented as vectors (see below for a formal definition). Our starting point is a GCN-VAE, and our first goal is a multitask reconstruction of both node features and network adjacency matrix. Then, as a second goal, we tune the architecture to be able to scale the number of embedding dimensions dedicated to feature-only reconstruction, adjacency-only reconstruction, or shared feature-adjacency information, while keeping the number of trainable parameters in the model constant.

4.3.1 Multitask graph convolutional autoencoder

We begin with the graph-convolutional variational autoencoder developed by [115], which stacks graph-convolutional (GC) layers [116] in the encoder part of a variational autoencoder [189, 112] to obtain a lower dimensional embedding of the input structure. This embedding is then used for the reconstruction of the original graph (and in our case, also of the features) in the decoding part of the model. Similarly to [116], we use two GC layers in our encoder and generate Gaussian-distributed node embeddings at the bottleneck layer of the autoencoder. We now introduce each phase of our embedding method in formal terms.

Encoder

We are given an undirected unweighted feautured graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $N = |\mathcal{V}|$ nodes, each node having a D-dimensional feature vector. Loosely following the notations of [115], we note $\mathbf{A}$ the graph's $N \times N$ adjacency matrix (diagonal elements set to 0), $\mathbf{X}$ the $N \times D$ matrix of node features, and $\mathbf{X}_i$ the D-dimensional feature vector of a node i.

The encoder part of our model is where F-dimensional node embeddings are generated. It computes $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, two $N \times F$ matrices, which parametrise a stochastic embedding of each node:

$$\boldsymbol{\mu} = \mathrm{GCN}_{\boldsymbol{\mu}}(\mathbf{X}, \mathbf{A}) \quad \text{and} \quad \log \boldsymbol{\sigma} = \mathrm{GCN}_{\boldsymbol{\sigma}}(\mathbf{X}, \mathbf{A}).$$

1. The implementation of our model is available online at github.com/ixxi-dante/an2vec.

Here we use two graph-convolutional layers for each parameter set, with shared weights at the first layer and parameter-specific weights at the second layer:

$$\mathrm{GCN}_\alpha(\mathbf{X}, \mathbf{A}) = \hat{\mathbf{A}}\,\mathrm{ReLU}(\hat{\mathbf{A}}\mathbf{X}\mathbf{W}_0^{enc})\mathbf{W}_{1,\alpha}^{enc}$$

In this equation, W_0^{enc} and $W_{1,\alpha}^{enc}$ are the weight matrices for the linear transformations of each layer's input; ReLU refers to a rectified linear unit [165]; and following the formalism introduced in [116], $\hat{\mathbf{A}}$ is the standard normalised adjacency matrix with added self-connections, defined as:

$$\begin{aligned}\hat{\mathbf{A}} &= \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\\ \tilde{\mathbf{A}} &= \mathbf{A} + \mathbf{I}_N\\ \tilde{D}_{ii} &= \sum_j \tilde{A}_{ij}\end{aligned}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix.

Embedding

The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ produced by the encoder define the distribution of an F-dimensional stochastic embedding $\boldsymbol{\xi}_i$ for each node i, defined as:

$$\boldsymbol{\xi}_i|\mathbf{A}, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2)).$$

Thus, for all the nodes we can write a probability density function over a given set of embeddings $\boldsymbol{\xi}$, in the form of an $N \times F$ matrix:

$$q(\boldsymbol{\xi}|\mathbf{X}, \mathbf{A}) = \prod_{i=1}^{N} q(\boldsymbol{\xi}_i|\mathbf{A}, \mathbf{X}).$$

Decoder

The decoder part of our model aims to reconstruct both the input node features and the input adjacency matrix by producing parameters of a generative model for each of the inputs. On one hand, the adjacency matrix $\mathbf{A}$ is modelled as a set of independent Bernoulli random variables, whose parameters come from a bi-linear form applied to the output of a single dense layer:

$$\begin{aligned}A_{ij}|\boldsymbol{\xi}_i, \boldsymbol{\xi}_j &\sim \mathrm{Ber}(\mathrm{MLB}(\boldsymbol{\xi})_{ij})\\ \mathrm{MLB}(\boldsymbol{\xi}) &= \mathrm{sigmoid}(\boldsymbol{\gamma}^T\mathbf{W}_{\mathbf{A},1}^{dec}\boldsymbol{\gamma})\\ \boldsymbol{\gamma} &= \mathrm{ReLU}(\boldsymbol{\xi}\mathbf{W}_{\mathbf{A},0}^{dec}).\text{‘}\end{aligned}$$

Similarly to above, $W^{dec}_{\mathbf{A},0}$ is the weight matrix for the first adjacency matrix decoder layer, and $W^{dec}_{\mathbf{A},1}$ is the weight matrix for the bilinear form which follows.

On the other hand, features can be modelled in a variety of ways, depending on whether they are binary or continuous, and if their norm is constrained or not. Features in our experiments are one-hot encodings, so we model the reconstruction of the feature matrix $\mathbf{X}$ by using N single-draw D-categories multinomial random variables. The parameters of those multinomial variables are computed from the embeddings with a two-layer perceptron: [2]

$$\begin{aligned}\mathbf{X}_i|\boldsymbol{\xi}_i &\sim \text{Multinomial}(1, \text{MLP}(\boldsymbol{\xi})_i)\\ \text{MLP}(\boldsymbol{\xi}) &= \text{softmax}(\text{ReLU}(\boldsymbol{\xi}\mathbf{W}^{dec}_{\mathbf{X},0})\mathbf{W}^{dec}_{\mathbf{X},1})\end{aligned}$$

In the above equations, $\text{sigmoid}(z) = \frac{1}{1+e^{-z}}$ refers to the logistic function applied element-wise on vectors or matrices, and $\text{softmax}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$ refers to the normalised exponential function, also applied element-wise, with j running along the rows of matrices (and along the indices of vectors).

Thus we can write the probability density for a given reconstruction as:

$$\begin{aligned}p(\mathbf{X}, \mathbf{A}|\boldsymbol{\xi}) &= p(\mathbf{A}|\boldsymbol{\xi})p(\mathbf{X}|\boldsymbol{\xi})\\ p(\mathbf{A}|\boldsymbol{\xi}) &= \prod_{i,j=1}^{N} \text{MLB}(\boldsymbol{\xi})_{ij}^{A_{ij}}(1-\text{MLB}(\boldsymbol{\xi})_{ij})^{1-A_{ij}}\\ p(\mathbf{X}|\boldsymbol{\xi}) &= \prod_{i=1}^{N}\prod_{j=1}^{D} \text{MLP}(\boldsymbol{\xi})_{ij}^{X_{ij}}\end{aligned}$$

Learning

The variational autoencoder is trained by minimising an upper bound to the marginal likelihood-based loss [189] defined as:

$$\begin{aligned}-\log p(\mathbf{A}, \mathbf{X}) &\leq \mathcal{L}(\mathbf{A}, \mathbf{X})\\ &= D_{KL}(q(\boldsymbol{\xi}|\mathbf{A}, \mathbf{X})||\mathcal{N}(0, \mathbf{I}_F))\\ &\quad - \mathbb{E}_{q(\boldsymbol{\xi}|\mathbf{A},\mathbf{X})}[\log(p(\mathbf{A}, \mathbf{X}|\boldsymbol{\xi}, \boldsymbol{\theta})p(\boldsymbol{\theta}))]\\ &= \mathcal{L}_{KL} + \mathcal{L}_{\mathbf{A}} + \mathcal{L}_{\mathbf{X}} + \mathcal{L}_{\boldsymbol{\theta}}\end{aligned}$$

2. Other types of node features are modelled according to their constraints and domain. Binary features are modelled as independent Bernoulli random variables. Continuous-range features are modelled as Gaussian random variables in a similar way to the embeddings themselves.

Here $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence between the distribution of the embeddings and a Gaussian Prior, and $\boldsymbol{\theta}$ is the vector of decoder parameters whose associated loss $\mathcal{L}_{\boldsymbol{\theta}}$ acts as a regulariser for the decoder layers.[3] Computing the adjacency and feature reconstruction losses by using their exact formulas is computationally not tractable, and the standard practice is instead to estimate those losses by using an empirical mean. We generate K samples of the embeddings by using the distribution $q(\boldsymbol{\xi}|\mathbf{A}, \mathbf{X})$ given by the encoder, and average the losses of each of those samples[4] [189, 112]:

$$\begin{aligned}
\mathcal{L}_{\mathbf{A}} &= -\mathbb{E}_{q(\boldsymbol{\xi}|\mathbf{A},\mathbf{X})}[\log p(\mathbf{A}|\boldsymbol{\xi}, \boldsymbol{\theta})] \\
&\simeq -\frac{1}{K}\sum_{k=1}^{K}\sum_{i,j=1}^{N}\big[A_{ij}\log(\mathrm{MLB}(\boldsymbol{\xi}^{(k)})_{ij}) \\
&\qquad + (1-A_{ij})\log(1-\mathrm{MLB}(\boldsymbol{\xi}^{(k)})_{ij})\big] \\
\mathcal{L}_{\mathbf{X}} &= -\mathbb{E}_{q(\boldsymbol{\xi}|\mathbf{A},\mathbf{X})}[\log p(\mathbf{X}|\boldsymbol{\xi}, \boldsymbol{\theta})] \\
&\simeq -\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{D} X_{ij}\log(\mathrm{MLP}(\boldsymbol{\xi}^{(\boldsymbol{k})})_{ij})
\end{aligned}$$

Finally, for diagonal Gaussian embeddings such as the ones we use, $\mathcal{L}_{KL}$ can be expressed directly [112]:

$$\mathcal{L}_{KL} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{F}\mu_{ij}^2 + \sigma_{ij}^2 - 2\log\sigma_{ij} - 1$$

Loss adjustments

In practice, to obtain useful results a few adjustments are necessary to this loss function. First, given the high sparsity of real-world graphs, the A_{ij} and $1-A_{ij}$ terms in the adjacency loss must be scaled respectively up and down in order to avoid globally near-zero link reconstruction probabilities. Instead of penalising reconstruction proportionally to the overall number of errors in edge prediction, we want false negatives (A_{ij} terms) and false positives ($1 - A_{ij}$ terms) to contribute equally to the reconstruction loss, independent of graph sparsity. Formally, let $d = \frac{\sum_{ij} A_{ij}}{N^2}$ denote the density of the graph's adjacency matrix ($d = \frac{N-1}{N} \times \mathrm{density}(\mathcal{G})$); then we replace

3. Indeed, following [189] we assume $\boldsymbol{\theta} \sim \mathcal{N}(0, \kappa_{\boldsymbol{\theta}}\mathbf{I})$, such that $\mathcal{L}_{\boldsymbol{\theta}} = -\log p(\boldsymbol{\theta}) = \frac{1}{2}\dim(\boldsymbol{\theta})\log(2\pi\kappa_{\boldsymbol{\theta}}) + \frac{1}{2\kappa_{\boldsymbol{\theta}}}||\boldsymbol{\theta}||_2^2$.

4. In practice, $K = 1$ is often enough.

$\mathcal{L}_{\mathbf{A}}$ by the following re-scaled estimated loss (the so-called "balanced cross-entropy"):

$$\tilde{\mathcal{L}}_{\mathbf{A}} = -\frac{1}{K}\sum_{k=1}^{K}\sum_{i,j=1}^{N}\frac{1}{2}\left[\frac{A_{ij}}{d}\log(\mathrm{MLB}(\boldsymbol{\xi}^{(k)})_{ij}) + \frac{1-A_{ij}}{1-d}\log(1-\mathrm{MLB}(\boldsymbol{\xi}^{(k)})_{ij})\right]$$

Second, we correct each component loss for its change of scale when the shapes of the inputs and the model parameters change: $\mathcal{L}_{KL}$ is linear in N and F, $\tilde{\mathcal{L}}_{\mathbf{A}}$ is quadratic in N, and $\mathcal{L}_{\mathbf{X}}$ is linear in N (but not in F, remember that $\sum_j X_{ij} = 1$ since each $\mathbf{X}_i$ is a single-draw multinomial).

Beyond dimension scaling, we also wish to keep the values of $\tilde{\mathcal{L}}_{\mathbf{A}}$ and $\mathcal{L}_{\mathbf{X}}$ comparable and, doing so, maintain a certain balance between the difficulty of each task. As a first approximation to the solution, and in order to avoid more elaborate schemes which would increase the complexity of our architecture (such as [37]), we divide both loss components by their values at maximum uncertainty [5], respectively $\log 2$ and $\log D$.

Finally, we make sure that the regulariser terms in the loss do not overpower the actual learning terms (which are now down-scaled close to 1) by adjusting $\kappa_{\boldsymbol{\theta}}$ and an additional factor, κ_{KL}, which scales the Kullback-Leibler term. [6] These adjustments lead us to the final total loss the model is trained for:

$$\mathcal{L} = \frac{\tilde{\mathcal{L}}_{\mathbf{A}}}{N^2\log 2} + \frac{\mathcal{L}_{\mathbf{X}}}{N\log D} + \frac{\mathcal{L}_{KL}}{NF\kappa_{KL}} + \frac{||\boldsymbol{\theta}||_2^2}{2\kappa_{\boldsymbol{\theta}}}$$

where we have removed constant terms with respect to trainable model parameters.

4.3.2 Scaling shared information allocation

The model we just presented uses all dimensions of the embeddings indiscriminately to reconstruct the adjacency matrix and the node features. While this can be useful in some cases, it cannot adapt to different interdependencies between graph structure and node features; in cases where the two are not strongly correlated, the embeddings would lose information by conflating features and graph structure. Therefore our second aim is to adjust the dimensions of the embeddings used exclusively for feature reconstruction, or for adjacency reconstruction, or used for both.

In a first step, we restrict which part of a node's embedding is used for each task. Let $F_{\mathbf{A}}$ be the number of embedding dimensions we will allocate to adjacency ma-

5. That is, $p(A_{ij}|\boldsymbol{\xi},\boldsymbol{\theta}) = \frac{1}{2} \quad \forall i,j$, and $p(X_{ij}|\boldsymbol{\xi},\boldsymbol{\theta}) = \frac{1}{D} \quad \forall i,j$.
6. We use $\kappa_{KL} = 2\kappa_{\boldsymbol{\theta}} = 10^3$.

trix reconstruction only, $F_{\mathbf{X}}$ the number of dimensions allocated to feature reconstruction only, and $F_{\mathbf{AX}}$ the number of dimensions allocated to both. We have $F_{\mathbf{A}} + F_{\mathbf{AX}} + F_{\mathbf{X}} = F$. We further introduce the following notation for the restriction of the embedding of node i to a set of dedicated dimensions $\{a, \ldots, b\}$ [7]:

$$\boldsymbol{\xi}_{i,a:b} = (\xi_{ij})_{j \in \{a,\ldots,b\}}$$

This extends to the full matrix of embeddings similarly:

$$\boldsymbol{\xi}_{a:b} = (\xi_{ij})_{i \in \{1,\ldots,N\}, j \in \{a,\ldots,b\}}$$

Using these notations we adapt the decoder to reconstruct adjacency and features as follows:

$$\begin{aligned} A_{ij} | \boldsymbol{\xi}_{i,1:F_{\mathbf{A}}+F_{\mathbf{AX}}}, \boldsymbol{\xi}_{j,1:F_{\mathbf{A}}+F_{\mathbf{AX}}} &\sim \mathrm{Ber}(\mathrm{MLB}(\boldsymbol{\xi}_{1:F_{\mathbf{A}}+F_{\mathbf{AX}}})_{ij}) \\ \mathbf{X}_i | \boldsymbol{\xi}_{i,F_{\mathbf{A}}+1:F} &\sim \mathrm{Multinomial}(1, \mathrm{MLP}(\boldsymbol{\xi}_{F_{\mathbf{A}}+1:F})_i) \end{aligned}$$

In other words, adjacency matrix reconstruction relies on $F_{\mathbf{A}} + F_{\mathbf{AX}}$ embedding dimensions, feature reconstruction relies on $F_{\mathbf{X}} + F_{\mathbf{AX}}$ dimensions, and $F_{\mathbf{AX}}$ overlapping dimensions are shared between the two. Our reasoning is that for datasets where the dependency between features and network structure is strong, shallow models with higher overlap value will perform better than models with the same total embedding dimensions F and less overlap, or will perform on par with models that have more total embedding dimensions and less overlap. Indeed, the overlapping model should be able to extract the information shared between features and network structure and store it in the overlapping dimensions, while keeping the feature-specific and structure-specific information in their respective embedding dimensions. This is to compare to the non-overlapping case, where shared network-feature information is stored redundantly, both in feature- and structure-specific embeddings, at the expense of a larger number of distinct dimensions.

Therefore, to evaluate the performance gains of this architecture, one of our measures is to compare the final loss for different hyperparameter sets, keeping $F_{\mathbf{A}} + F_{\mathbf{AX}}$ and $F_{\mathbf{X}} + F_{\mathbf{AX}}$ fixed and varying the overlap size $F_{\mathbf{AX}}$. Now, to make sure the training losses for different hyperparameter sets are comparable, we must maintain the overall number of trainable parameters in the model fixed. The decoder already has a constant number of trainable parameters, since it only depends on the number of dimensions used for decoding features ($F_{\mathbf{X}} + F_{\mathbf{AX}}$) and adjacency matrix ($F_{\mathbf{A}} + F_{\mathbf{AX}}$), which are themselves fixed.

7. Note that the order of the indices does not change the training results, as the model has no notion of ordering inside its layers. What follows is valid for any permutation of the dimensions, and the actual indices only matter to downstream interpretation of the embeddings after training.

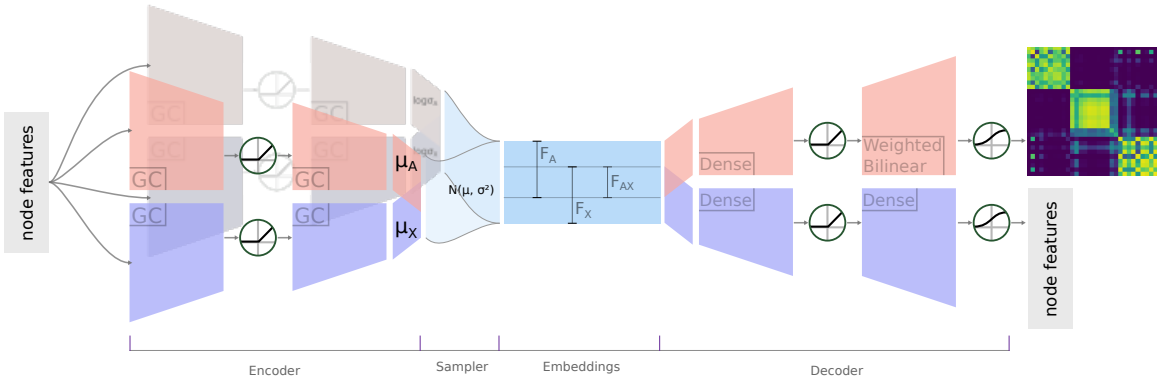


Figure 4-1: Diagram of the overlapping embedding model we propose. Red and blue blocks with a layer name (GC, Dense, Weighted Bilinear) indicate actual layers, with their activation function depicted to the right as a curve in a green circle (either ReLU or sigmoid). Red blocks concern processing for the adjacency matrix, blue blocks processing for the node features. The encoder is made of four parallel GC pipelines producing $\mu_{\mathbf{A}}$, $\mu_{\mathbf{X}}$, $\log \sigma_{\mathbf{A}}$ and $\log \sigma_{\mathbf{X}}$ (the last two being grayed out in the background). Their output is then combined to create the overlap, then used by the sampler to create the node embeddings. The decoder processes parts of the node embeddings and separately reconstructs the adjacency matrix (top) and the node features (bottom).

On the other hand, the encoder requires an additional change. We maintain the dimensions of the encoder-generated $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ parameters fixed at $F_{\mathbf{A}} + 2F_{\mathbf{AX}} + F_{\mathbf{X}}$ (independently from $F_{\mathbf{AX}}$, given the constraints above), and reduce those outputs to $F_{\mathbf{A}} + F_{\mathbf{AX}} + F_{\mathbf{X}}$ dimensions by averaging dimensions $\{F_{\mathbf{A}} + 1, \ldots, F_{\mathbf{A}} + F_{\mathbf{AX}}\}$ and $\{F_{\mathbf{A}} + F_{\mathbf{AX}} + 1, \ldots, F_{\mathbf{A}} + 2F_{\mathbf{AX}}\}$ together.[8] In turn, this model maintains a constant number of trainable parameters, while allowing us to adjust the number of dimensions $F_{\mathbf{AX}}$ shared by feature and adjacency reconstruction (keeping $F_{\mathbf{A}} + F_{\mathbf{AX}}$ and $F_{\mathbf{X}} + F_{\mathbf{AX}}$ constant). Figure 4-1 schematically represents this architecture.

8. Formally:

$$\begin{aligned}\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_{1:F_{\mathbf{A}}} \,\|\, &\frac{1}{2}(\boldsymbol{\mu}_{F_{\mathbf{A}}+1:F_{\mathbf{A}}+F_{\mathbf{AX}}} + \boldsymbol{\mu}_{F_{\mathbf{A}}+F_{\mathbf{AX}}+1:F_{\mathbf{A}}+2F_{\mathbf{AX}}}) \\ &\| \, \boldsymbol{\mu}_{F_{\mathbf{A}}+2F_{\mathbf{AX}}+1:F_{\mathbf{A}}+2F_{\mathbf{AX}}+F_{\mathbf{X}}} \\ \log \tilde{\boldsymbol{\sigma}} = \log \boldsymbol{\sigma}_{1:F_{\mathbf{A}}} \,\|\, &\frac{1}{2}(\log \boldsymbol{\sigma}_{F_{\mathbf{A}}+1:F_{\mathbf{A}}+F_{\mathbf{AX}}} + \log \boldsymbol{\sigma}_{F_{\mathbf{A}}+F_{\mathbf{AX}}+1:F_{\mathbf{A}}+2F_{\mathbf{AX}}}) \\ &\| \, \log \boldsymbol{\sigma}_{F_{\mathbf{A}}+2F_{\mathbf{AX}}+1:F_{\mathbf{A}}+2F_{\mathbf{AX}}+F_{\mathbf{X}}}\end{aligned}$$

where $\|$ denotes concatenation along the columns of the matrices.

4.4 Results

We are interested in measuring two main effects: first, the variation in model performance as we increase the overlap in the embeddings, and second, the capacity of the embeddings with overlap (versus no overlap) to capture and benefit from dependencies between graph structure and node features. To that end, we train overlapping and non-overlapping models on synthetic data with different degrees of correlation between network structure and node features.

4.4.1 Synthetic featured networks

We use a Stochastic Block Model [92] to generate synthetic featured networks, each with M communities of $n = 10$ nodes, with intra-cluster connection probabilities of 0.25, and with inter-cluster connection probabilities of 0.01. Each node is initially assigned a colour which encodes its feature community; we shuffle the colours of a fraction $1 - \alpha$ of the nodes, randomly sampled. This procedure maintains constant the overall count of each colour, and lets us control the correlation between the graph structure and node features by moving α from 0 (no correlation) to 1 (full correlation).

Node features are represented by a one-hot encoding of their colour (therefore, in all our scenarios, the node features have dimension $M = N/n$). However, since in this case all the nodes inside a community have exactly the same feature value, the model can have difficulties differentiating nodes from one another. We therefore add a small Gaussian noise ($\sigma = .1$) to make sure that nodes in the same community can be distinguished from one another.

Note that the feature matrix has less degrees of freedom than the adjacency matrix in this setup, a fact that will be reflected in the plots below. However, opting for this minimal generative model lets us avoid the parameter exploration of more complex schemes for feature generation, while still demonstrating the effectiveness of our model.

4.4.2 Comparison setup

To evaluate the efficiency of our model in terms of capturing meaningful correlations between network and features, we compare overlapping and non-overlapping models as follows. For a given maximum number of embedding dimensions F_{max}, the overlapping models keep constant the number of dimensions used for adjacency matrix reconstruction and the number of dimensions used for feature reconstruction, with the same amount allocated to each task: $F^{ov}_{\mathbf{A}} + F^{ov}_{\mathbf{AX}} = F^{ov}_{\mathbf{X}} + F^{ov}_{\mathbf{AX}} = \frac{1}{2}F_{max}$. However they vary the overlap $F^{ov}_{\mathbf{AX}}$ from 0 to $\frac{1}{2}F_{max}$ by steps of 2. Thus the total

number of embedding dimensions F varies from F_{max} to $\frac{1}{2}F_{max}$, and as F decreases, $F^{ov}_{\mathbf{AX}}$ increases. We call one such model $\mathcal{M}^{ov}_F$.

Now for a given overlapping model $\mathcal{M}^{ov}_F$, we define a reference model $\mathcal{M}^{ref}_F$, which has the same total number of embedding dimensions, but without overlap: $F^{ref}_{\mathbf{AX}} = 0$, and $F^{ref}_{\mathbf{A}} = F^{ref}_{\mathbf{X}} = \frac{1}{2}F$ (explaining why we vary F with steps of 2). Note that while the reference model has the same information bottleneck as the overlapping model, it has less trainable parameters in the decoder, since $F^{ref}_{\mathbf{A}} + F^{ref}_{\mathbf{AX}} = F^{ref}_{\mathbf{X}} + F^{ref}_{\mathbf{AX}} = \frac{1}{2}F$ will decrease as F decreases. Nevertheless, this will not be a problem for our measures, since we will be mainly looking at the behaviour of a given model for different values of α (i.e. the feature-network correlation parameter).

For our calculations (if not noted otherwise) we use synthetic networks of $N = 1000$ nodes (i.e. 100 clusters), and set the maximum embedding dimensions F_{max} to 20. For all models, we set the intermediate layer in the encoder and the two intermediate layers in the decoder to an output dimension of 50, and the internal number of samples for loss estimation at $K = 5$. We train our models for 1000 epochs using the Adam optimiser [111] with a learning rate of 0.01 (following [115]), after initialising weights following [67]. For each combination of F and α, the training of the overlapping and reference models is repeated 20 times on independent featured networks.

Since the size of our synthetic data is constant, and we average training results over independently sampled data sets, we can meaningfully compare the averaged training losses of models with different parameters. We therefore take the average best training loss of a model to be our main measure, indicating the capacity to reconstruct an input data set for a given information bottleneck and embedding overlap.

4.4.3 Advantages of overlap

Absolute loss values

Figure 4-2 shows the variation of the best training loss (total loss, adjacency reconstruction loss, and feature reconstruction loss) for both overlapping and reference models, with α ranging from 0 to 1 and F decreasing from 20 to 10 by steps of 2. One curve in these plots represents the variation in losses of a model with fixed F for data sets with increasing correlation between network and features; each point aggregates 20 independent trainings, used to bootstrap 95% confidence intervals.

We first see that all losses, whether for overlapping model or reference, decrease as we move from the uncorrelated scenario to the correlated scenario. This is true despite the fact that the total loss is dominated by the adjacency reconstruction loss,

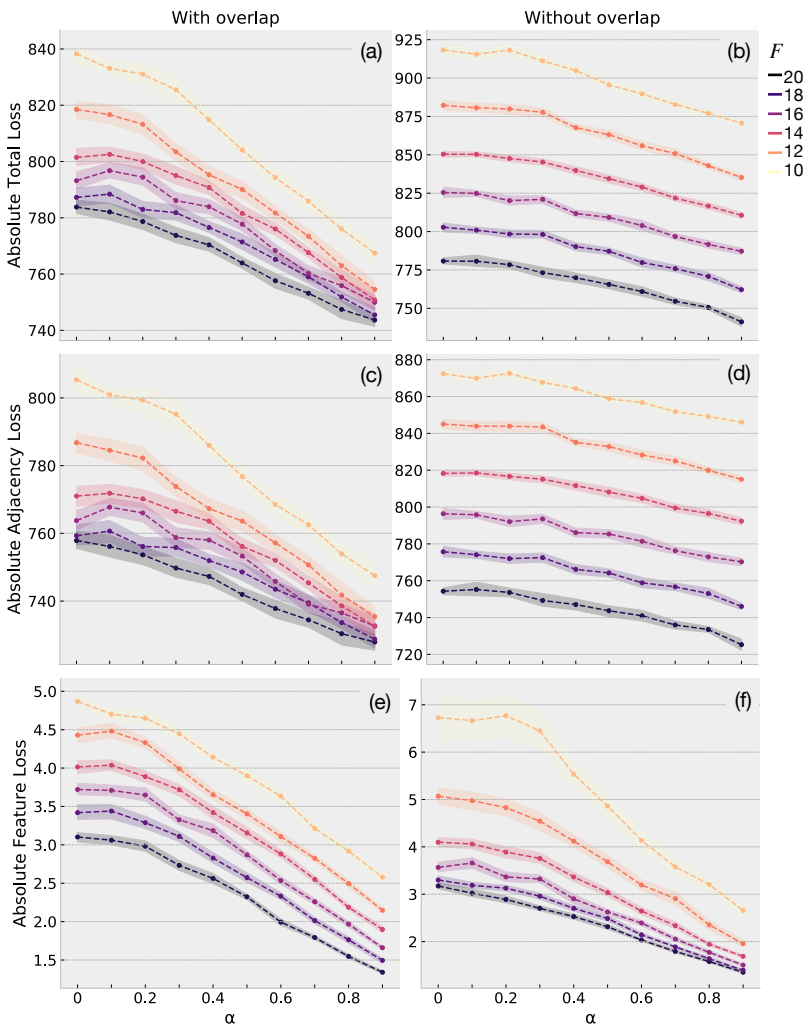


Figure 4-2: Absolute training loss values of overlapping and reference models. The curve colours represents the total embedding dimensions F, and the x axis corresponds to feature-network correlation. The top row is the total loss, the middle row is the adjacency matrix reconstruction loss and the bottom row is the feature reconstruction loss. The left column shows overlapping models, and the right column shows reference non-overlapping models.

as feature reconstruction is an easier task overall. Second, recall that the decoder in a reference model has less parameters than its corresponding overlapping model of the same F dimensions (except for zero overlap), such that the reference is less powerful and produces higher training losses. The absolute values of the losses for overlap and reference models are therefore not directly comparable. However, the changes in slopes are meaningful. Indeed, we note that the curve slopes are steeper for models with higher overlap (lower F) than for lower overlap (higher F), whereas they seem relatively independent for the reference models of different F. In other words, as we increase the overlap, our models seem to benefit more from an increase

in network-feature correlation than what a reference model benefits.

Relative loss disadvantage

In order to assess this trend more reliably, we examine losses relative to the maximum embedding models. Figure 4-3 plots the loss disadvantage that overlap and reference models have compared to their corresponding model with $F = F_{max}$, that is, $\frac{\mathcal{L}_{\mathcal{M}_F} - \mathcal{L}_{\mathcal{M}_{F_{max}}}}{\mathcal{L}_{\mathcal{M}_{F_{max}}}}$. We call this the *relative loss disadvantage* of a model. In this plot, the height of a curve thus represents the magnitude of the decrease in performance of a model $\mathcal{M}_F^{ov|ref}$ relative to the model with maximum embedding size, $\mathcal{M}_{F_{max}}^{ov|ref}$. Note that for both the overlap model and the reference model, moving along one of the curves does not change the number of trainable parameters in the model.

As the correlation between network and features increases, we see that the relative loss disadvantage decreases in overlap models, and that the effect is stronger for higher overlaps. In other words, when the network and features are correlated, the overlap captures this joint information and compensates for the lower total number of dimensions (compared to $\mathcal{M}_{F_{max}}^{ov|ref}$): the model achieves a better performance than when network and features are more independent. Strikingly, for the reference model these curves are flat, thus indicating no variation in relative loss disadvantage with varying network-feature correlations in these cases. This confirms that the new measure successfully controls for the baseline decrease of absolute loss values when the network-features correlation increases, as observed in Figure 4-2. Our architecture is therefore capable of capturing and taking advantage of some of the correlation by leveraging the overlap dimensions of the embeddings.

Finally note that for high overlaps, the feature reconstruction loss value actually increases a little when α grows. The behaviour is consistent with the fact that the total loss is dominated by the adjacency matrix loss (the hardest task). In this case it seems that the total loss is improved more by exploiting the gain of optimising for adjacency matrix reconstruction, and paying the small cost of a lesser feature reconstruction, than decreasing both adjacency matrix and feature losses together. If wanted, this strategy could be controlled using a gradient normalisation scheme such as [37].

4.4.4 Standard benchmarks

Finally we compare the performance of our architecture to other well-known embedding methods, namely spectral clustering (SC) [215], DeepWalk (DW) [176], the vanilla non-variational and variational Graph Auto-Encoders (GAE and VGAE) [115], and GraphSAGE [80] which we look at in more detail. We do so on two tasks: (i) the

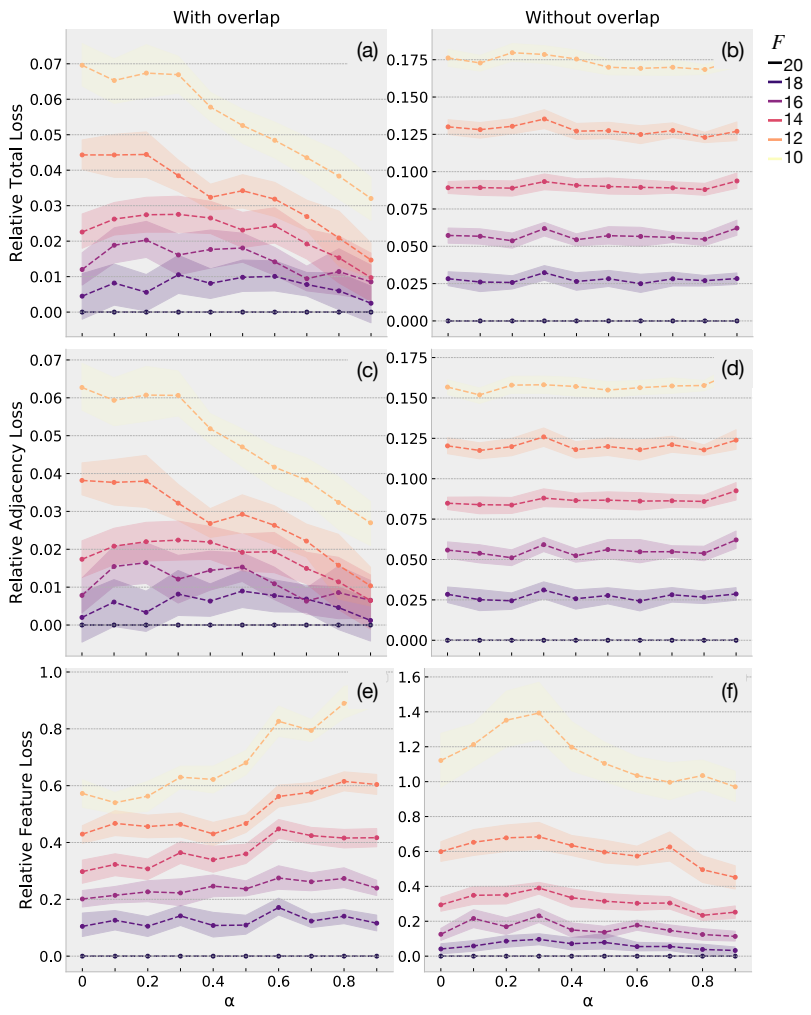


Figure 4-3: Relative loss disadvantage for overlapping and reference models. The curve colours represents the total embedding dimensions F, and the x axis corresponds to feature-network correlation. The top row is the total loss, the middle row is the adjacency matrix reconstruction loss and the bottom row is the feature reconstruction loss. The left column shows overlapping models, and the right column shows reference non-overlapping models. See main text for a discussion.

link prediction task introduced by [115] and (ii) a node classification task, both on the Cora, CiteSeer and PubMed datasets, which are regularly used as citation network benchmarks in the literature [201, 166]. Note that neither SC nor DW support feature information as an input.

The Cora and CiteSeer datasets are citation networks made of respectively 2708 and 3312 machine learning articles, each assigned to a small number of document classes (7 for Cora, 6 for CiteSeer), with a bag-of-words feature vector for each article (respectively 1433 and 3703 words). The PubMed network is made of 19717 diabetes-related

articles from the PubMed database, each assigned to one of three classes, with article feature vectors containing *term frequency-inverse document frequency* (TF/IDF) scores for 500 words.

Link prediction

The link prediction task consists in training a model on a version of the datasets where part of the edges has been removed, while node features are left intact. A test set is formed by randomly sampling 15% of the edges combined with the same number of random disconnected pairs (non-edges). Subsequently the model is trained on the remaining dataset where 15% of the real edges are missing.

Method	**Cora**		**CiteSeer**		**PubMed**	
	AUC	AP	AUC	AP	AUC	AP
SC	84.6 ± 0.01	88.5 ± 0.00	80.5 ± 0.01	85.0 ± 0.01	84.2 ± 0.02	87.8 ± 0.01
DW	83.1 ± 0.01	85.0 ± 0.00	80.5 ± 0.02	83.6 ± 0.01	84.4 ± 0.00	84.1 ± 0.00
GAE	91.0 ± 0.02	92.0 ± 0.03	89.5 ± 0.04	89.9 ± 0.05	**96.4** ± 0.00	**96.5** ± 0.00
VGAE	91.4 ± 0.01	92.6 ± 0.01	90.8 ± 0.02	92.0 ± 0.02	94.4 ± 0.02	94.7 ± 0.02
AN2VEC-0	89.5 ± 0.01	90.6 ± 0.01	91.2 ± 0.01	91.5 ± 0.02	91.8 ± 0.01	93.2 ± 0.01
AN2VEC-16	89.4 ± 0.01	90.2 ± 0.01	91.1 ± 0.01	91.3 ± 0.02	92.1 ± 0.01	92.8 ± 0.01
AN2VEC-S-0	92.9 ± 0.01	93.4 ± 0.01	94.3 ± 0.01	94.8 ± 0.01	95.1 ± 0.01	95.4 ± 0.01
AN2VEC-S-16	**93.0** ± 0.01	**93.5** ± 0.00	**94.9** ± 0.00	**95.1** ± 0.00	93.1 ± 0.01	93.1 ± 0.01

Table 4.1: Link prediction task in citation networks. SC, DW, GAE and VGAE values are from [115]. Error values indicate the sample standard deviation.

We pick hyperparameters such that the restriction of our model to VGAE would match the hyperparameters used by [115]. That is a 32-dimensions intermediate layer in the encoder and the two intermediate layers in the decoder, and 16 embedding dimensions for each reconstruction task ($F_{\mathbf{A}} + F_{\mathbf{AX}} = F_{\mathbf{X}} + F_{\mathbf{AX}} = 16$). We call the zero-overlap and the full-overlap versions of this model AN2VEC-0 and AN2VEC-16 respectively. In addition, we test a variant of these models with a shallow adjacency matrix decoder, consisting of a direct inner product between node embeddings, while keeping the two dense layers for feature decoding. Formally: $A_{ij}|\boldsymbol{\xi}_i, \boldsymbol{\xi}_j \sim \text{Ber}(\boldsymbol{\xi}_i^T \boldsymbol{\xi}_j)$. This modified overlapping architecture can be seen as simply adding the feature decoding and embedding overlap mechanics to the vanilla VGAE. Consistently, we call the zero-overlap and full-overlap versions AN2VEC-S-0 and AN2VEC-S-16.

We follow the test procedure laid out by [115]: we train for 200 epochs using the Adam optimiser [111] with a learning rate of .01, initialise weights following [67], and repeat each condition 10 times. The $\boldsymbol{\mu}$ parameter of each node's embedding is then used for link prediction (i.e. the parameter is put through the decoder directly without sampling), for which we report *area under the ROC curve* and *average precision* scores in Table 4.1. [9]

9. Note that in [115], the training set is also 85% of the full dataset, and test and validation sets

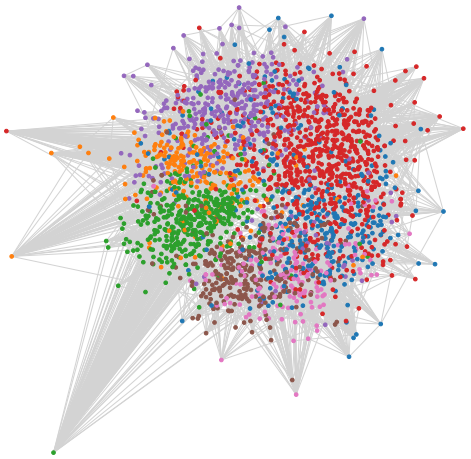


Figure 4-4: Cora embeddings created by AN2VEC-S-16, downscaled to 2D using Multidimensional scaling. Node colours correspond to document classes, and network links are in grey.

We argue that AN2VEC-0 and AN2VEC-16 should have somewhat poorer performance than VGAE. These models are required to reconstruct an additional output, which is not directly used to the link prediction task at hand. First results confirmed our intuition. However, we found that the shallow decoder models AN2VEC-S-0 and AN2VEC-S-16 perform consistently better than the vanilla VGAE for Cora and CiteSeer while their deep counterparts (AN2VEC-0 and AN2VEC-16) outperforms VGAE for all datasets. As neither AN2VEC-0 nor AN2VEC-16 exhibited over-fitting, this behaviour is surprising and calls for further explorations which are beyond the scope of this work (in particular, this may be specific to the link prediction task). Nonetheless, the higher performance of AN2VEC-S-0 and AN2VEC-S-16 over the vanilla VGAE on Cora and CiteSeer confirms that including feature reconstruction in the constraints of node embeddings is capable of increasing link prediction performance when feature and structure are not independent (consistent with [61, 203, 222]). An illustration of the embeddings produced by AN2VEC-S-16 on Cora is shown in Figure 4-4.

On the other hand, performance of AN2VEC-S-0 on PubMed is comparable with GAE and VGAE, while AN2VEC-S-16 has slightly lower performance. The fact that lower overlap models perform better on this dataset indicates that features and structure are less congruent here than in Cora or CiteSeer (again consistent with the comparisons found in [222]). Despite this, an advantage of the embeddings produced by the AN2VEC-S-16 model is that they encode *both* the network structure and the

are formed with the remaining edges, respectively 10% and 5% (and the same amount of non-edges). Here, since we use the same hyperparameters as [115] we do not need a validation set. We therefore chose to use the full 15% remaining edges (with added non-edges) as a test set, as explained above.

node features, and can therefore be used for downstream tasks involving both types of information.

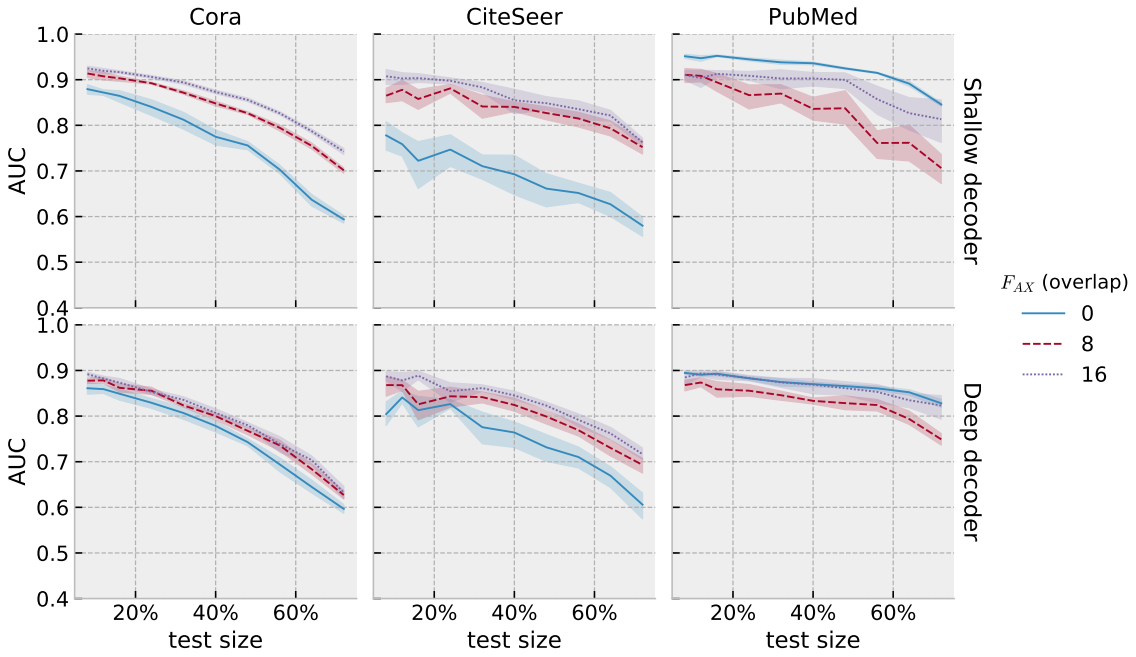
We further explore the behaviour of the model for different sizes of the training set, ranging from 10% to 90% of the edges in each dataset (reducing the training set accordingly), and compare the behaviour of AN2VEC to GraphSAGE. To make the comparison meaningful we train two variants of the two-layer GraphSAGE model with mean aggregators and no bias vectors: one with an intermediate layer of 32 dimensions and an embedding layer of 16 dimensions (roughly equivalent in dimensions to the full overlap AN2VEC models), the second with an intermediate layer of 64 dimensions and an embedding layer of 32 dimensions (roughly equivalent to no overlap in AN2VEC). Both layers use neighbourhood sampling, 10 neighbours for the first layer and 5 for the second. Similarly to the shallow AN2VEC decoder, each pair of node embeddings is reduced by inner product and a sigmoid activation, yielding a scalar prediction between 0 and 1 for each possible edge. The model is trained on minibatches of 50 edges and non-edges (edges generated with random walks of length 5), learning rate 0.001, and 4 total epochs. Note that on Cora, one epoch represents about 542 minibatches, [10] such that 4 epochs represent about 2166 gradient updates; thus with a learning rate of 0.001, we remain comparable to the 200 full batches with learning rate 0.01 used to train AN2VEC.

Figure 4-5 plots the AUC produced by AN2VEC and GraphSAGE for different training set sizes and different embedding sizes (and overlaps, for AN2VEC), for each dataset. As expected, the performance of both models decreases as the size of the test set increases, though less so for AN2VEC. For Cora and CiteSeer, similarly to Table 4.1, higher overlaps and a shallow decoder in AN2VEC give better performance. Notably, the shallow decoder version of AN2VEC with full overlap is still around .75 for a test size of 90%, whereas both GraphSAGE variants are well below .65. For PubMed, as in Table 4.1, the behaviour is different to the first two datasets, as overlaps 0 and 16 yield the best results. As for Cora and CiteSeer, the approach taken by AN2VEC gives good results: with a test size of 90%, all AN2VEC deep decoder variants are still above .75 (and shallow decoders above .70), whereas both GraphSAGE variants are below .50.

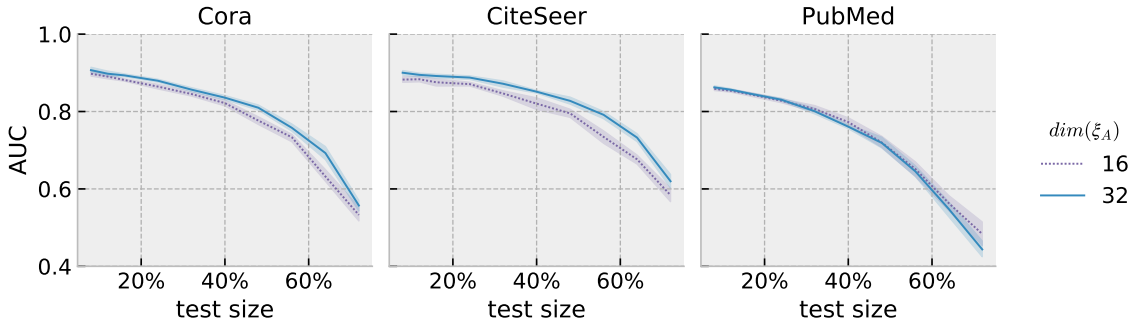
Node classification

Since the embeddings produced also to encode feature information, we then evaluate the model's performance on a node classification task. Here the models are trained on a version of the dataset where a portion of the nodes (randomly selected) have

10. One epoch is 2708 nodes $\times$ 5 edges per node $\times$ 2 (for non-edges) = 27080 training edges or non-edges; divided by 50, this makes 541.6 minibatches per epoch.



(a) AN2VEC



(b) GraphSAGE

Figure 4-5: AUC for link prediction using AN2VEC and GraphSAGE over all datasets. AN2VEC top row is the shallow decoder variant, and the bottom row is the deep decoder variant; colour and line styles indicate different levels of overlap. GraphSAGE colours and line styles indicate embedding size as described in the main text (colour and style correspond to the comparable variant of AN2VEC). Each point on a curve aggregates 10 independent training runs.

been removed; next, a logistic classifier [11] is trained on the embeddings to classify training nodes into their classes; finally, embeddings are produced for the removed nodes, for which we show the F1 scores of the classifier.

Figure 4-6 shows the results for AN2VEC and GraphSAGE on all datasets. The

11. Using Scikit-learn's [174] interface to the liblinear library, with one-vs-rest classes.

scale of the reduction in performance as the test size increases is similar for both models (and similar to the behaviour for link prediction), though overlap and shallow versus deep decoding seem to have less effect. Still, the deep decoder is less affected by the change in test size than the shallow decoder; and contrary to the link prediction case, the 0 overlap models perform best (on all datasets). Overall, the performance levels of GraphSAGE and AN2VEC on this task are quite similar, with slightly better results of AN2VEC on Cora, slightly stronger performance for GraphSAGE on CiteSeer, and mixed behaviour on PubMed (AN2VEC is better for small test sizes and worse for large test sizes).

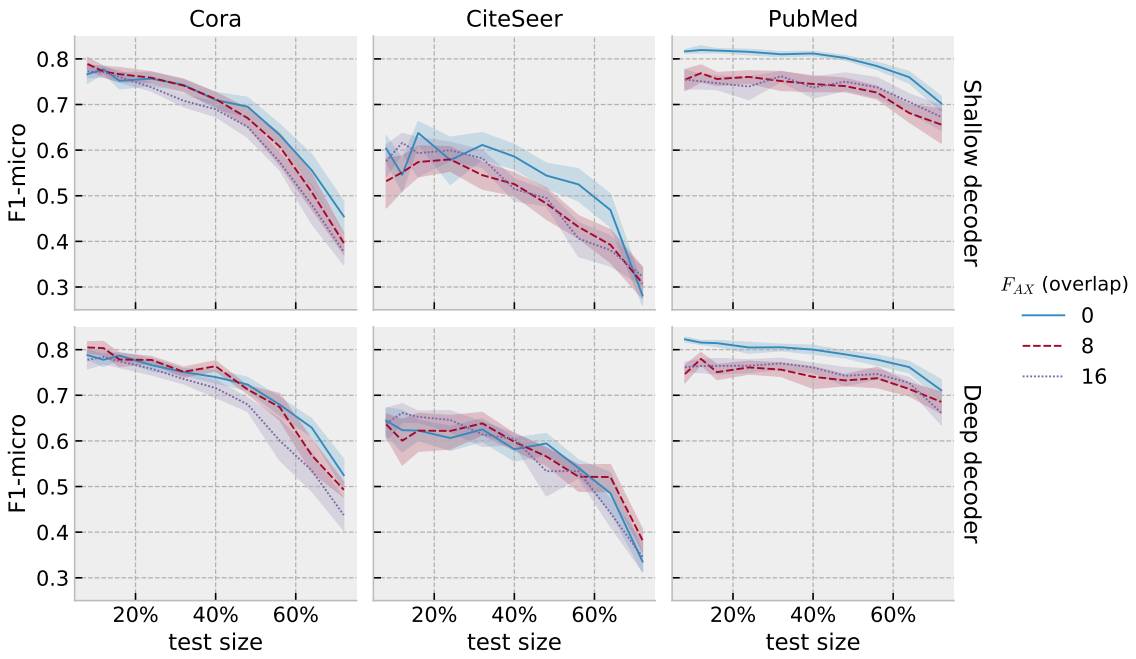
Variable embedding size

Finally, we explore the behaviour of AN2VEC for different embedding sizes. We train models with $F_{\mathbf{A}} = F_{\mathbf{X}} \in \{8, 16, 24, 32\}$ and overlaps 0, 8, 16, 24, 32 (whenever there are enough dimensions to do so), with variable test size. Figure 4-7 shows the AUC scores for link prediction, and Figure 4-8 shows the F1-micro scores for node classification, both on CiteSeer (the behaviour is similar on Cora, though less salient). For link prediction, beyond confirming trends already observed previously, we see that models with less total embedding dimensions perform slightly better than models with more total dimensions. More interestingly, all models seem to reach a plateau at overlap 8, and then exhibit a slightly fluctuating behaviour as overlap continues to increase (in models that have enough dimensions to do so). This is valid for all test sizes, and suggests (i) that at most 8 dimensions are necessary to capture the commonalities between network and features in CiteSeer, and (ii) that having more dimensions to capture either shared or non-shared information is not necessarily useful. In other words, 8 overlapping dimensions seem to capture most of what can be captured by AN2VEC on the CiteSeer dataset, and further increase in dimensions (either overlapping or not) would capture redundant information.

Node classification, on the other hand, does not exhibit any consistent behaviour beyond the reduction in performance as the test size increases. Models with less total dimensions seems to perform slightly better at 0 overlap (though this behaviour is reversed on Cora), but neither the ordering of models by total dimensions nor the effect of increasing overlap are consistent across all conditions. This suggests, similarly to Figure 4-6a, that overlap is less relevant to this particular node classification scheme than it is to link prediction.

4.5 Conclusions

In this chapter, we proposed an attributed network embedding method based on the combination of Graph Convolutional Networks and Variational Autoencoders.



(a) AN2VEC

(b) GraphSAGE

Figure 4-6: F1-micro score for node classification using AN2VEC and GraphSAGE over all datasets. AN2VEC top row is the shallow decoder variant, and the bottom row is the deep decoder variant; colour and line styles indicate different levels of overlap. GraphSAGE colours and line styles indicate embedding size as described in the main text (colour and style correspond to the comparable variant of AN2VEC). Each point on a curve aggregates 10 independent training runs.

Beyond the novelty of this architecture, it is able to consider jointly network information and node attributes for the embedding of nodes. We further introduced a control parameter able to regulate the amount of information allocated to the reconstruction of the network, the features, or both. In doing so, we showed how shallow versions of the proposed model outperform the corresponding non-interacting reference embeddings on given benchmarks, and demonstrated how this overlap parameter consistently captures joint network-feature information when they are correlated.

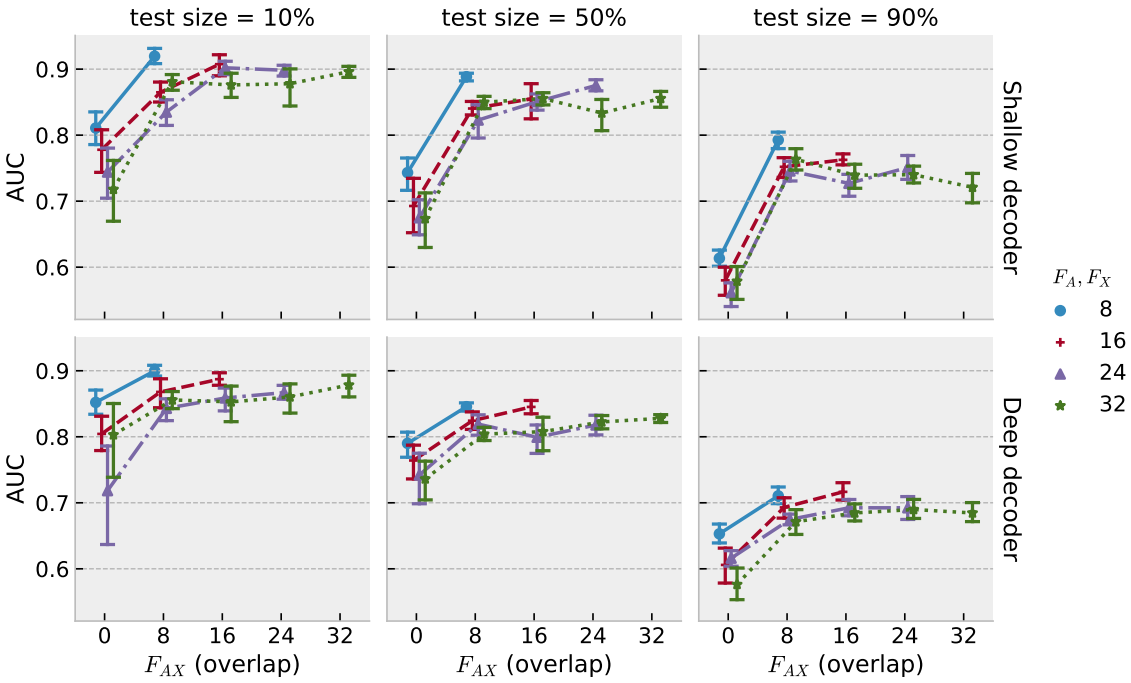


Figure 4-7: AUC for link prediction using AN2VEC on CiteSeer, as a function of overlap, with variable total embedding dimensions. Columns correspond to different test set sizes. Top row is with shallow decoder, bottom row with deep decoder. Colours, as well as marker and line styles, indicate the number of embedding dimensions available for adjacency and features.

Our method therefore opens several new lines of research and applications in fields where attributed networks are relevant. Examples range from the coevolution of social networks to dynamical processes taking place on them such as language or trending information, and can be potentially envisaged in systems biology or social sciences. In particular, it is well-poised by design to shed light upon the hidden network and individual factors that drive socioeconomic inequalities by providing us with an edge in terms of predictive performance and indicating which similarities, structural, feature-related, or both, better explain why a a certain individual is of a given socioeconomic status.

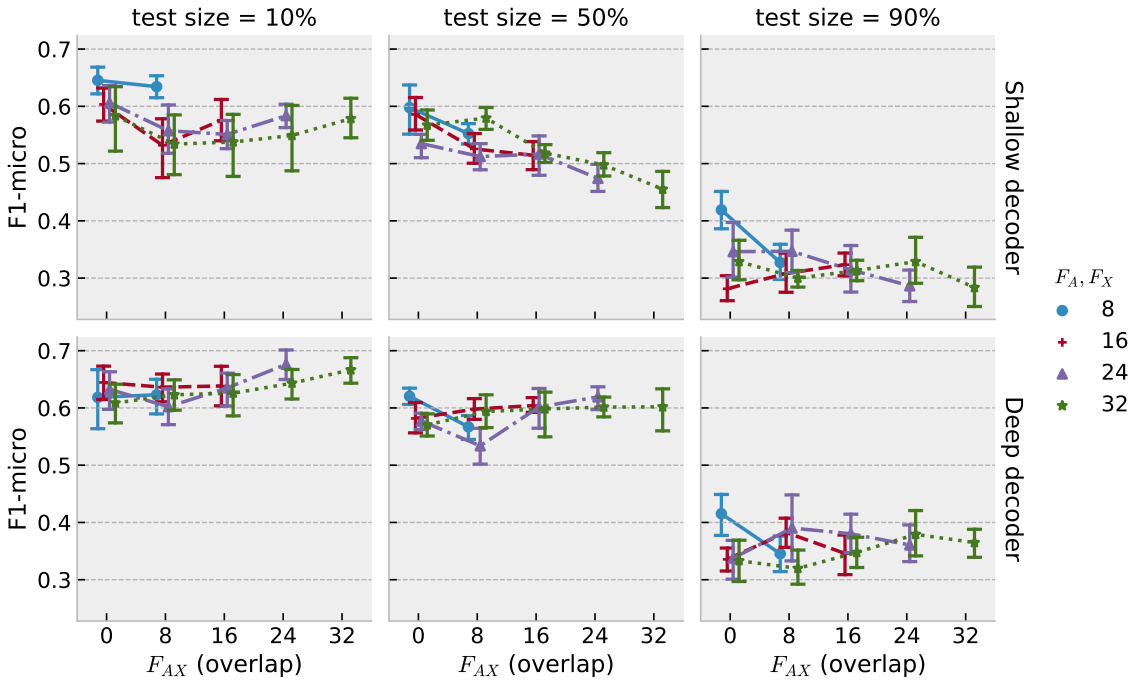


Figure 4-8: F1-micro score for node classification using AN2VEC on CiteSeer, as a function of overlap, with variable total embedding dimensions. Columns correspond to different test set sizes. Top row is with shallow decoder, bottom row with deep decoder. Colours, as well as marker and line styles, indicate the number of embedding dimensions available for adjacency and features.

Chapter 5

Deep learning captured socioeconomic correlations of urban patterns

This work is currently in submission as a journal paper in Scientific Reports [?] with myself as leading author

5.1 Introduction

Up to this point, all the proposed models were assumed to be operating in data-rich environments in which individuals could be observed from a sociological perspective based on whom they were connected to as well as on the syntactic style and semantic content of their tweets. Because of this, their ability to have a meaningful societal impact is limited by the heavy data reliance needed to train them. Most of the world is nevertheless living in a rather data scarce environment, as social media penetration rate can be dramatically lower in developing countries than in developed ones and census information can be severely outdated [20] thus failing to provide the updated upscaled picture of society that would be desirable. Because of this, satellite and aerial data have become the cornerstone of new models aimed at curbing this scarcity of information as the surge in works fueled by them testifies. These models generally rely on Convolutional Neural Network based models which not only require vast amounts of data to be trained but have also been shown to generalize poorly among different countries [108]. In this chapter, we introduce a model able to generate socioeconomic predictions from aerial views at an unprecented spatial scale in France and interpret the model's predictions in terms of urban classes. Our hope in doing so is to provide a benchmark for future remote sensing socioeconomic predictions as well as to inform future works aimed at increasing the transferability of the aforementioned models, so that one day, models trained in data rich environments could

be used off the shelf in different data scarce ones.

Cities have become the economic bedrock of modern nations; in little more than a century they have gone from concentrating 13% to an estimated 55% of the world population with 600 of them currently accounting for around half of the global economic output ($34 trillion of GDP) [168]. This transition will likely be accelerated in the coming years as an estimated three billion people will move into cities by 2030. This increased urbanization is in turn a key driver of development as cities provide the national platform for shared prosperity. Indeed, the concentration of people in cities generate agglomeration economies where the sheer population density increases the ease of moving goods, people, and ideas by removing the physical spaces between people and firms and thus increasing the returns of urban proximity [66].

Nevertheless, while urbanization can entail economic dynamism, and social development, it can also create enormous social challenges, from the management of natural hazards and pollution to the exclusion of the poor from the city's socioeconomic fabric and the surge of social and economic inequalities. This last issue is especially acute in some of the cities with greatest concentration of wealth. For instance, Gourevitch et al. reported a 24 years difference in average lifespan in San Francisco for citizens inhabiting neighbouring census tracks [72], while recent statistics show the area of Greater Paris containing respectively 2 and 5 of the top 5 poorest and richest census tracks in all of France [105]. Providing a solution to these challenges is of paramount importance to fulfill the economic and social promises that cities hold and to avoid them becoming sources of social and political instability.

The successful development and deployment of urban solutions to address these issues requires however both detailed socioeconomic information at the city level as well as a fine-grained understanding on how the distribution of wealth and the underlying urban topology of the city are entangled. In order to generate the former, different types of proxies have been used such as communication patterns in call detail records [20] and social media [149] or restaurant data [50]. Most notably, some works have relied heavily upon large collections of satellite and street imagery to train deep learning models able to predict wealth either from visual features within each image (e.g Suel et al. for 3 UK cities [210]) or, in order to enhance the interpretability of the learned models, by predicting features known to correlate with wealth (e.g night time light intensities [108], models of car per census track [63]). Surprisingly, despite the overall use and reliability of these methods, no previous work has looked consistently at the features learned by these models, nor tried to uncover any existing correlations between the existing urban topology of the city and the high resolution socioeconomic map the model is tasked with predicting. Understanding these correlations is a milestone in ensuring the usability of these methods for policy makers, as they not only

provide an enhanced understanding on the model's attentions but they also might yield valuable insights for urban development and planning of economically deprived areas within the city.

Our aim is to close this gap by exploring how the weights of a Convolutional Neural Network (CNN) model trained to predict the socioeconomic status of a given location from its aerial image relate to the land use classes of that urban area. We first overlay three publicly available datasets to provide a complete description of 5 French cities in terms of socioeconomic and land use data as well as aerial imagery. We then train a deep learning model to predict accurately the socioeconomic status of inhabited 200m x 200m tiles and backtrack the gradients to generate high resolution class discriminative activation maps. These are later projected onto the original image and overlayed with the land use data to generate empirical statistics on the features used by each model to predict socioeconomic status in terms of land use classes. We find that this framework enables the inference of socioeconomic status at scales rarely seen before, as well the observation of distinct city to city patterns of correlations between urban topology and the spread of wealth. More precisely, we find that when inferring socioeconomic status, our CNN models disregard existing correlations between land use and socioeconomic data and focus mostly on features contained within residential areas to draw their predictions. In what follows we will cover the datasets this work is reliant upon to then cover the architecture of the CNN used to predict income from satellite imagery as well as the activation maps enabling the examination of the features models drew their predictions from. We will eventually examine how urban structure and income are related to each other and how our deep learning model relies on one to predict the other.

5.2 Datasets

Socioeconomic Data. We obtained detailed sociodemographic information from the French National Institute of Statistics and Economic Studies (INSEE) 2019 gridded dataset [106]. This data corpus describes the winsorized average household income, estimated from the 2015 tax return in France, for each 4 hectare (200m x 200m) square patch across the whole French territory (Figure 5-1c). In what follows these socioeconomic patches will also be referred as census cells. Contrary to previous versions of this dataset, the winsorization is done at the department level, which enables us to get an in-depth view of socioeconomic disparities for the whole country. For each city, income values were partitioned into one of five ($n_{SES} = 5$) socioeconomic classes defined by the five quantiles of the city-wise income distribution so that classes 0 and 4 correspond respectively to the bottom and top 20% of earners in each city. For the remainder of this study, we singled out 5 major cities

located within the French metropolitan territory, namely: **Paris**, **Lyon**, **Marseille**, **Nice** and **Lille**. Geographical boundaries for each of these cities/urban areas were obtained from the ensuing Land Use Data. The corresponding spatial distribution of income may be observed in Figures 5-5a, A-1a, A-2a, A-3a, A-4a where each coloured pixel corresponds to a single socioeconomic patch.

Aerial Imagery Data. We obtained high resolution orthophotography taken between 2013 and 2016 of the complete French metropolitan territory (20cm/pixel) from the National Geographical Information Institute (IGN). This dataset is provided as a series of georeferenced 5km x 5km tiles at the department (administrative delimitation) level as shown in Figure 5-1a. In order to extract the aerial imagery corresponding to each socioeconomic square patch (aerial tiles in what follows), we derived the following procedure: First, for each tile T, we identified all the socioeconomic patches contained exclusively within T and we extracted them. Then, we identified all the socioeconomic cells overlapping multiple aerial tiles. These tiles are subsequently merged and the corresponding patches extracted. The results of this procedure can be seen in Figure 5-3a where we show the image associated to a given socioeconomic patch associated with a particular socioeconomic label. Contrary to previous works [210, 63, 108], we chose not to rely on Google Maps/Streets as a source of data for two basic reasons. First, API crawled data lacks georeferencing which hinders any attempt of overlaying activation maps with land use data. Second, these datasets have recently become harder to access [217] at the scale needed for conducting this type of work, leading us to eventually discard them. So far, our dataset consists of a series of 200m x 200m aerial tiles fully covering five French cities and each associated to one of five socioeconomic classes.

Land Use Data. We gathered land use information from the 2012 European Union Urban Atlas dataset. This dataset provides high-resolution maps of roughly 700 urban areas of more than 100,000 inhabitants in EU28 and EFTA countries. In doing so, it yields consistent land cover maps encoded via detailed polygons organized in commonly used ESRI shapefiles and covering 27 standardized land use classes (see Figure 5-1b). This number was later reduced to 19 for this study as infrequent and similar classes were respectively discarded and merged (see explicit list in Table 5.1). Further information regarding the procedure by which this dataset was generated can be obtained from the European Environment Agency [202] and is explained in the Supplementary Information (SI). The land cover information is hence overlayed on top of the image of every socioeconomic patch, yielding results shown in Figure 5-3d.

Note that all the collected datasets can be openly accessed and emanate from a single 4-year time window hence reducing temporal misalignment between them to a minimum.

Figure 5-1: Sample of Overlayed Datasets (Paris): **(a)** 5km x 5km aerial tile (20cm/pixel), **(b)** Land cover map of the same area, each color representing a different urban class (not shown here for clarity), **(c)** Spatial Distribution of Income: each patch corresponds to a single 200m x 200m patch with precise income data.

5.3 Methods

5.3.1 Model development and training

To estimate the socioeconomic status (SES) of every image in our dataset, we trained a modified EfficientNetB0 [214] (EB0, see Figure 5-2) model on each urban area to predict the SES label of a census cell from its aerial tile. This model was chosen over more classical ones (VGGs, ResNets) mostly for its architectural design: by making use of an effective compound coefficient to scale up CNNs , EfficientNets have been shown to achieve much better accuracy and efficiency than other ConvNets enabling us to make use of the original fine-grained resolution of the aerial imagery to predict the SES at a lower computational cost.

Our model used the original EB0 architecture consisting of serial mobile inverted bottleneck blocks (MBConv) augmented with squeeze-and-excitation optimization, to which we added a max pooling layer, to downsample the feature maps, followed by global average pooling and dense layers (see Figure 5-2). In order for the model to yield an estimated probability for the input to belong to each of the quantiles, we followed the approach used by Suel et al. [210], and connected the dense layer to a binomial layer taking a single p value, interpreted as a probability with which Bernoulli trials are performed, as input and outputting the probabilities for each SES class $Y = [\hat{y}_1, ..., \hat{y}_{n_{SES}}]$, with:

$$\hat{y}_k = \binom{n_{SES}}{k} p^{k-1}(1-p)^{n_{SES}-k}, \quad k \in [1, ..., n_{SES}]$$

In doing so, the model becomes sensitive to the ordinal relationship existing between quantiles. Training is then performed with a classic categorical cross-entropy loss.

It is worth mentioning that the original EB0 model was trained on object-centric 224 x 224 x 3 images and therefore the weights the model was initialized with were unfit for the dataset we wanted to apply it unto. Because of this, even though the CNN's weights were initially set to those of a model trained on ImageNet (except for the final added layers which are initialized randomly), we didn't freeze any of the model's layers to train the model from scratch. This of course assumes the optimization procedure will more readily converge by adapting the weights from the optimal ones learned at lower resolutions than by actually learning them from scratch, which stands to reason since the first layers of the network will already be sensitive to low-level features.

In all our experiments, we augmented the data through random horizontal and vertical flipping of the input images and the images were scaled to 800 x 800 x 3 so as not to decrease excessively the original image resolution whilst, at the same time, not overwhelming our computational resources. An Adam variant of Stochastic Gradient Descent was used to learn the models weights. Initially, we started with a learning rate of 8e-5 and decreased it by 90% every 3 epochs the validation loss sees no improvement. All networks were trained on either NVIDIA Tesla K20X or V100 graphics processing units for at most 30 epochs with 2500 samples in each epoch, stopping the training whenever the loss in the validation set didn't decrease for more than 10 epochs. Furthermore, within each city, we used five-fold cross validation. In each fold, 80% of data was used for training the network and the remaining 20% were withheld. The training set was further randomly subdivided with a 75-25% split for inner-fold training and validation. Every reported performance metrics is then averaged over all folds.

Figure 5-2: Model Architecture: The model takes the aerial tile as an input which is then fed through several MBConv blocks. The feature maps end up going through a global average pooling layer and a dense layer to output a single value p. From it, probabilities for each socioeconomic class are generated from a binomial distribution $B_{(N_{SES},p)}$

5.3.2 Guided Grad-CAM (GG-CAM)

Grad-CAM generates coarse, discriminative regions for each of the socioeconomical classes the models were trained to predict. It is calculated as the rectified linear of the weighted sum of the feature maps A^k from the last convolutional layer (`conv2d_65` in the original EfficientNetB0 model) of the CNN:

$$L^c{}_{\text{Grad-CAM}} = \text{ReLU}\Big(\sum_k \alpha_k^c A^k\Big)$$

This weighted combination of forward activation maps is based on the weights α_k^c, defined by :

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{i,j}^k}$$

where Z is a normalization constant, y^c is the prediction score for SES class c and $A_{i,j}^k$ is the ij^{th} element of A^k. Each α_k^c represents a partial linearization of the CNN downstream from A^k, and captures the importance of the feature map k for a SES class c. Once computed, activation maps are L^1-normalized.

Guided BackPropagation, on the other hand, is a method that captures non class-discriminative details of visual components that are of some importance to the network by suppressing the flow of gradients through neurons wherein either of input or incoming gradients were negative. It is computed as :

$$R^l = f^{l'}\ \text{ReLU}\ \Big(\frac{\partial f^{out}}{\partial f^{l'}}\Big)\frac{\partial f^{out}}{\partial f^{l'}}$$

with R_l guided backpropagation product of the l-th layer, f_l and f^{out} feature maps respectively of the $l-$th and last convolutional layer and $f^{l'} = \text{ReLU}\ (f^l)$. Guided GradCAM are then generated via the Hadamard product of GradCAM and the guided backpropagation saliency map. Contrary to previous works, the outcome of this operation isn't normalized any further so as to enable comparison between samples from the same city.

5.3.3 Guided gradient-weight class activation maps

Once all models were trained, we cross-examined the features in the original images that activated the model the most when predicting a given SES class. To do so, we applied a guided gradient-weighted class activation mapping [200] (Guided Grad-CAM) to identify salient visual features underlying the model's predictions. Previous uses of this method include the identification of patterns learned by CNN models trained to predict Alzheimer's disease [102, 216], lung cancer [93] as well as to classify wildlife [158] and plants disease [219]. As we saw previously, Guided Grad-CAM follows the CNN's gradient flow from individual output classes back onto the original

image tile to establish an activation map, highlighting the input features most relevant to each CNN prediction. To generate such mappings, we first computed saliency maps via guided backpropagation, producing a pixel-space gradient map of predicted class scores with respect to pixel intensities of the input image. This gradient visualization, though fine-grained is known not to be class-discriminative. It can however be combined with Grad-CAM via pointwise multiplication to generate Guided Grad-CAM (see Appendix for more information). Grad-CAM itself results from the dot product of the feature map of the last convolutional layer and the partial derivatives of predicted class scores with respect to the neurons in the last convolutional layer. All mappings were done in a cluster of 96 CPUs.

We now define the setting we worked with. In what follows, we treated Guided-GradCAM maps as saliency maps describing pixels that were most determinant for the model's prediction. For a given city, we named S the set of aerial tiles corresponding to this city, F_s the CNN trained on that city (intaking a sample $s \in S$ and outputting its predicted class $\hat{y}_s \in [1...n_{SES}]$), N_{ua} the total number of different urban classes included in our dataset and A^s_{ij} the activation value of the Guided GradCAM for pixel (i, j) of tile s. Each tile is moreover defined by a mapping f linking each pixel to the urban class of the polygon it belongs to, so that, $\forall (i, j),\ f(i, j) \in [1..N_{ua}]$. We then introduce the total activation for urban class $u \in [1...N_{ua}]$ in tile $s \in S$ as :

$$A^s_u = \sum_{(i,j)|f(i,j)=u} A_{i,j}$$

It is worth noting then that each sample $s \in S$ is defined by the tuple $(A^s_u)_{u\in[1..N_{ua}]}$ where $A^s_u = \text{NA}$ for all urban classes u not contained in tile s. We then sought to probe the model's learned features by assessing the validity of the two following hypotheses:

— **H0**: All urban classes contain features equally contributing to the model's prediction for the top and bottom SES class.

— **H1**: Urban class activation maps are pairwise independent, i.e, activations for a given urban class are on average invariant to activation values for other urban classes.

In order to examine the former, we introduced the activation ratio for urban class u defined as the ratio between the sum of activations and the expected sum of activations in a random diffusion model concentrated by polygons of urban class u:

$$r^s_u = \frac{A^s_u}{F_u \sum_{k=1}^{N_{ua}} A^s_k} \text{ with } F_u \text{ fraction of area occupied by polygons of urban class } u$$

This metric actually describes how the pixels that most trigger the model's prediction are distributed across urban classes. For instance, if no urban class contains more

important features than others (**H0**) we should expect its expected value to be equal to 1 over all samples and for all urban classes contained in each sample. To quantify this as well as the distributions of activation values over different urban classes, we respectively introduce the expected activation ratio for urban class u, x_u and the co-activation ratio $x_{u,v}$ for pairs of urban classes u and v defined by:

$$x_u^\sigma = \mathbb{E}_\sigma(r_u^s) \text{ and } x_{u,v}^\sigma = \frac{\mathbb{E}_{\sigma'}(r_u^s)}{\mathbb{E}_\sigma(r_u^s)} - 1 \text{ with } \sigma' = \{s \in \sigma | A_u^s, A_v^s \neq \text{NA}\}$$

with $\sigma \subseteq S$, set of aerial tiles over which the expectation is computed. Notice that, in general, the co-activation ratio is not symmetric and should be close to 1 under **(H1)**.

Both metrics above are to be computed over a specific set of tiles for each city. To yield a better understanding of the features learned by each model when predicting wealth and poverty, we respectively merged the top two and bottom two SES classes that from here on out will be referred to as high SES and low SES. An overview of this process can be seen in Figure 5-3. The expected activation ratio and co-activation ratio for urban classes were then computed both in samples for which the model predicted low SES ($\sigma_{\text{low}} = \{s \in S | \hat{y}_s \in \{0, 1\}\}$) and high SES ($\sigma_{\text{high}} = \{s \in S | \hat{y}_s \in \{3, 4\}\}$ respectively).We finally define the expected activation ratio for low SES and high SES as:

$$\begin{cases} c_{0,1} = (x_u^{\sigma_{\text{low}}})_u \text{ and } c_{3,4} = (x_u^{\sigma_{\text{high}}})_u \\ d_{0,1} = (x_{u,v}^{\sigma_{\text{low}}})_{u,v} \text{ and } d_{3,4} = (x_{u,v}^{\sigma_{\text{high}}})_{u,v} \end{cases}$$

We have hitherto focused on the activation maps of each model to investigate the correlations learned between urban patterns and SES. To provide a more grounded comparison between the former and the actual link between SES and urban classes, we introduce the empirical probabilities $\hat{p}_{0,1} = (\hat{p}_{0,1}^u)_u$ and $\hat{p}_{3,4} = (\hat{p}_{3,4}^u)_u$ that a sample is of respectively low (high) SES given that it contains urban class u as well as the co-appearance gains:

$$\begin{cases} \hat{g}_{0,1} = (\hat{g}_{0,1}^{(u,v)})_{(u,v)} = (\hat{p}_{0,1}^{(u,v)} / \hat{p}_{0,1}^u - 1)_{(u,v)} \\ \hat{g}_{3,4} = (\hat{g}_{3,4}^{(u,v)})_{(u,v)} = (\hat{p}_{3,4}^{(u,v)} / \hat{p}_{3,4}^u - 1)_{(u,v)} \end{cases}$$

which are thought of as the increase in the likelihood of a sample being of respectively low (high) SES given that it contains urban class u co-appearing with urban class v. Notice as well that for a pair of urban classes (u, v) co-appearance gains are not symmetrical.

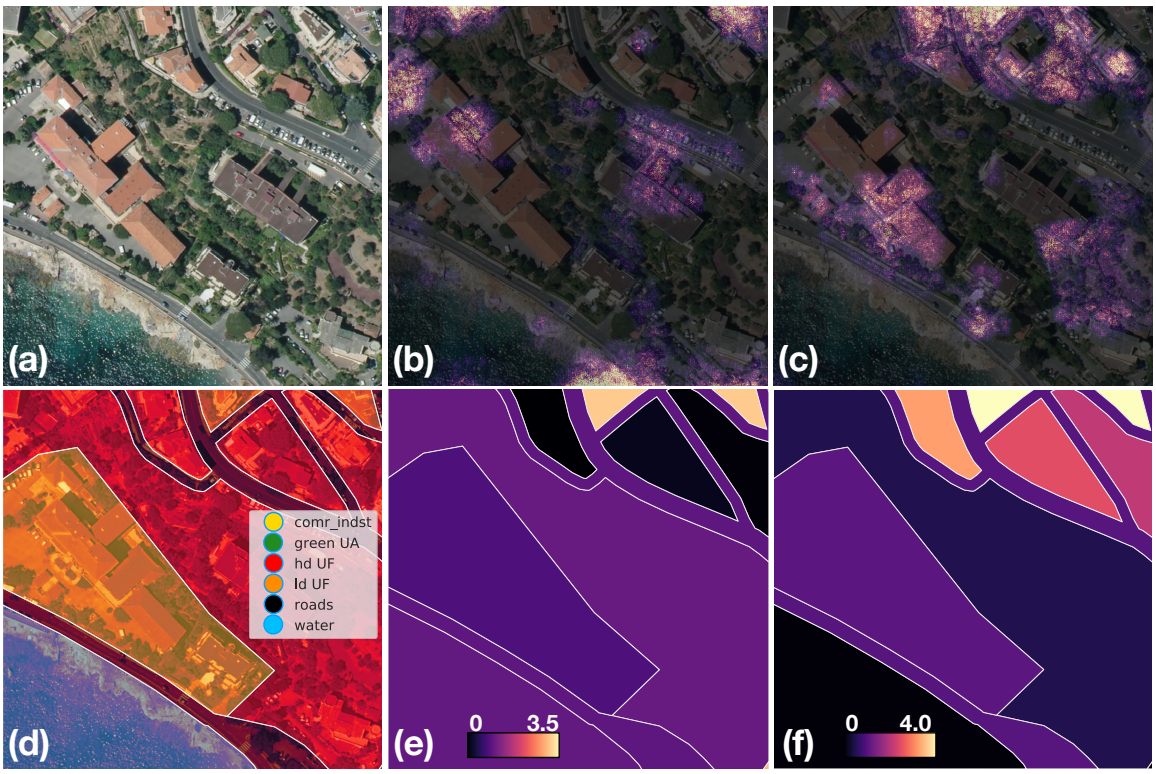


Figure 5-3: Model interpretability studies using Guided Grad-CAM (GGC). From an aerial tile (**a**), GGC is used to compute activation maps for the poorest (**b**) and wealthiest (**c**) socioeconomical class. The activation maps are then overlayed with the tile's tesselation into urban classes polygon (**d**) to compute the normalized ratio of activations per polygon for the poorest (**e**) and wealthiest (**f**) class.

5.4 Results

5.4.1 CNN-based socioeconomic predictions

Modified EfficientNetB0 models were separately trained on 5 cities (Paris, Lyon, Marseille, Nice, Lille) to predict 5-class socioeconomic status from 200m x 200m aerial tiles. To assess the quality of the predictions made by each model, we computed the metrics for reported by Suel et al. [210] for each city in our dataset. We report the 5-fold averaged observed performance per city in Figure 5-4. All values are computed for samples not seen either during training or validation of the model.

The quality of the predictions tends to be quite homogeneous between cities with the average Pearson correlation between true and predicted quantiles varying between 0.65 for Lyon and 0.71 in Nice. Furthermore, in all cities, the mean average error between predicted and observed socioeconomical class tends to be bounded between 0.72 for Paris and 0.80 for Lyon which in turn explains why misclassifications occur between neighbouring quantiles. It is worth mentioning that all tiles were classified within ± 1 of their class with similar accuracies to those reported in previous studies

(ranging from 0.82 to 0.85) [210]. Also, it is interesting to observe that between cities the worst performing classes predicted by each model tends to vary from one another: In Paris and Lyon, the fourth quantile appears to be the hardest one to predict correctly, while the third and second one seem to be the prone to errors respectively in Marseille and Nice.

We further report the observed and predicted spread of wealth for each of the five

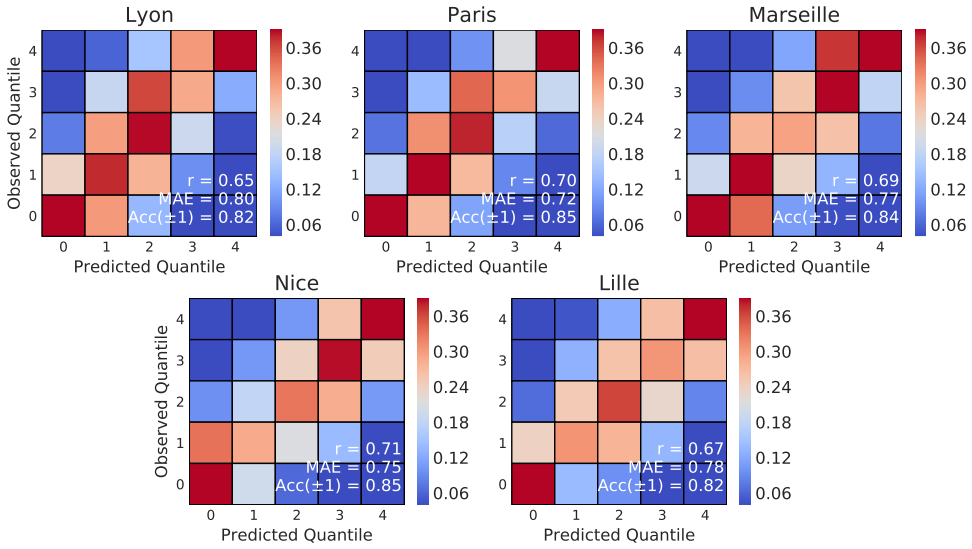


Figure 5-4: Observed performance of models trained to predict wealth in French cities: Confusion matrices between predicted and observed SES classes. Perfect prediction would correspond to a red diagonal and blue off-diagonal cells. In each plot, Pearson correlation coefficient ($r, p < 0.01$)) as well as mean average error (MAE) and accuracy within ± 1 classes are provided. All results are average over the k cross-validation folds (with $k = 5$).

cities we work with (see Paris in Figure 5-5 and others in Appendix A.1). Contrary to other studies [210, 63], each map pixel here represents a single sample corresponding to a unique tile, socioeconomic class pair. Consequently, predicted values are not aggregated but are individually computed. It is therefore interesting to see that neighbouring tiles of equal SES are overall predicted to be of the same socioeconomic class, hence recovering the small-scale spatial homogeneity of the original income distribution (see center of Paris, Lyon or south-west of Nice and Marseille). Moreover, conversely, regions characterized by socioeconomic disparity seem to be recovered as well in the prediction maps (see for instance in the west of Marseille, East of Lille, and downtown-suburbs predicted socioeconomic segregation in Paris), further showing that the trained models are able to recover the complex spatial distributions of income featured in our dataset.

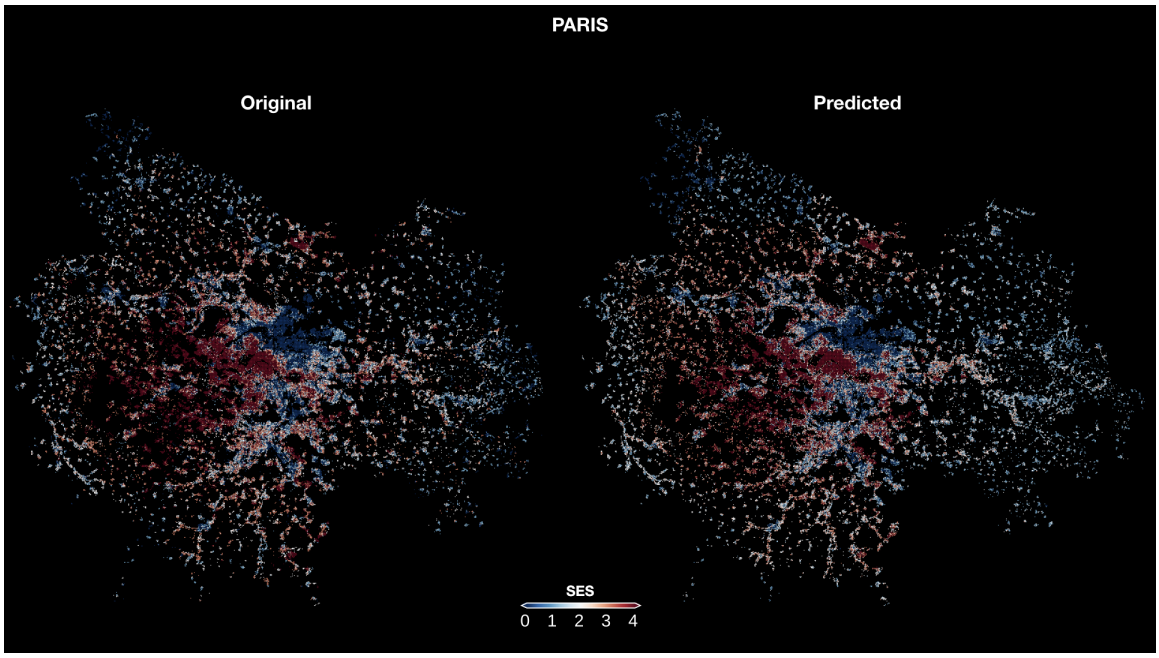


Figure 5-5: Maps of observed and predicted average income for Paris. Each pixel represents a single 200m x 200m tile which color codifies the average socioeconomic status of its inhabitants. Maps for the other cities can be found in the Appendix.

5.4.2 GradCAM derived model correlations between SES and urban topology

Univariate Correlations

We now sought to investigate the CNN model's internal logic for predicting the socioeconomic label of a given area. To do so, we applied guided gradient-weighted class activation mapping (Guided Grad-CAM) to identify the salient visual features underlying the model's predictions in terms of land use for samples predicted as low SES ($c_{0,1}$) and high SES($c_{3,4}$). As a means of comparison, we provide as well the empirical probabilities for a sample to actually be of low SES and high SES given the urban classes it contains ($\hat{p}_{0,1}$ and $\hat{p}_{3,4}$ respectively). Results for Paris can be found in Figure 5-6 while other cities are included in the SI.

When compared across different cities, the spectrum of activations per urban class differs from one city to the next, indicating the disparity of elements models trained on different settings focus on to predict socioeconomic status. In what follows, we will refer to a urban class u being overactivated if the expected activation ratio for that class is greater than one, i.e $x_u^\sigma > 1$, which is of course the expected value one would observe if activations were spread randomly across each each tile. Conversely, it is said to be underactivated whenever its expected activation ratio is below 1, i.e

$x_u^\sigma < 1$. To ease the description of the results, we grouped each urban class into four categories (see Figure 5-6 and Table 5.1).

ID	Abbreviation	Full Denomination
1	**agri_wetland**	Agricultural Areas/ Wetlands
2	**green ua**	Green Urban Areas
3	**ntral areas**	Natural Areas
4	**comr_indst**	Industrial, Commercial, Public, Military and Private Units
5	**leis fac.**	Sports and Leisure Facilities
6	**const/dmp**	Mine, Dump and Construction Sites
7	**no use**	Land without Current Use
8	**roads**	Other roads and associated land
9	**rlway**	Railways and Associated Land
10	**mtrways**	Fast Transit roads and associated land
11	**port**	Port Areas
12	**water**	Water Bodies
13	**osp/bch**	Open Spaces with little or no vegetation (beaches, dunes, bare rocks, glaciers)
14	**isoltd**	Isolated Structures
15	**vld uf**	Discontinuous Very Low Density Urban Fabric (s.l. $< 10\%$)
16	**ld uf**	Discontinuous Low Density Urban Fabric (s.l. 10% - 30%)
17	**md uf**	Discontinuous Medium Density Urban Fabric (s.l. 30% - 50%)
18	**hd uf**	Discontinuous Dense Urban Fabric (s.l. 50% - 80%)
19	**vhd uf**	Continuous Urban Fabric (s.l. $> 80\%$)

Table 5.1: Land cover classes (i.e urban classes) used in this study.

Looking back upon the spectrum of activations, certain structural elements tend to be shared across all trained models. Residential areas tend to be overall overactivated both when predicting low and high SES with some differences between cities: For models trained to predict SES in Lyon and Marseille, residential areas tend to activate more when predicting high SES rather than low SES (also to a lesser extent in Lille) while in Paris and Nice, on the contrary, residential areas appear overactivated when predicting low SES.

Infrastructural areas on the other hand follow quite a different pattern, as they tend to be underactivated in the prediction of high SES except for the city of Lille where both motorways and roads are triggered when predicting high SES. This trend is however not the case for low SES predictions, where railways and motorways are triggered in Lyon and Marseille while features contained within roads are generally

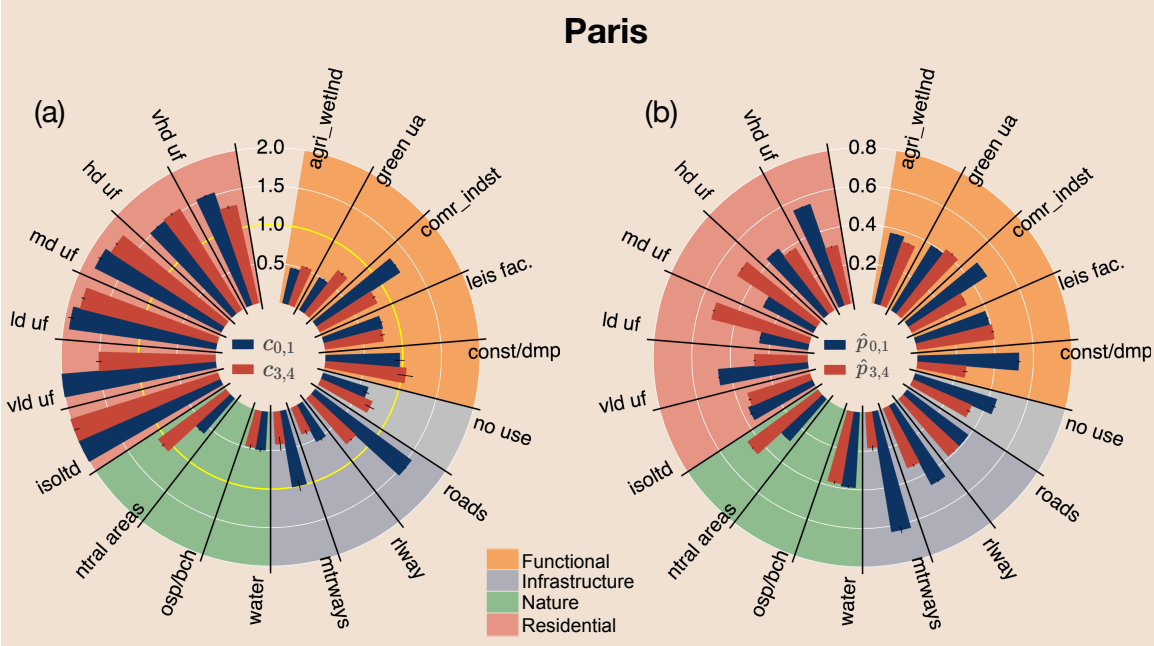


Figure 5-6: Correlations between urban topology and socioeconomic status in the city of Paris: (a) Mean model activation rate per urban class with 95% confidence interval for samples predicted as respectively as low SES (blue) or high SES (red) by the model (b) Estimated probability of a urban polygon belonging to the bottom or top quintile of the income distribution with bootstrapped 95% confidence interval.

active for all cities.
This is perhaps most surprising when we consider that the urban classes residential areas co-ocurr with are actually quite determinant in terms of the socioeconomic status of the sampl

When considering nature related areas, these tend to be ignored for high SES predictions while for low SES, this seems to be the case as well except for Nice and Marseille. A similar trend was observed for functional areas which tend to be under-activated as well for all SES predictions and all cities except commercial/industrial areas that trigger low SES predictions mostly. More generally, if we were to look at the above results as a whole, we would conclude that CNN models trained to predict SES in urban areas are mostly reliant on features lying within residential areas to draw their predictions, while the degree to which other amenities are determinant in the prediction shifts from city to city, therefore invalidating **(H0)**.

Furthermore, the spectrum of activation these models exhibit differs significantly from the urban classes that empirically determine SES the most, identified by means of the analysis of $\hat{p}_{0,1}$ and $\hat{p}_{3,4}$ (see Section 5.3.3). For instance, while the type of residential areas contained in a given tile greatly varies between high SES and low SES samples, with low and medium density areas being more prone to be of high

SES and very high density ones more likely to be of low SES, similar variations are observed as well in other categories: Tiles containing motorways, railways and commercial/industrial units are mostly associated to poor areas where as those including natural areas are more likely to be of high SES.

Bivariate Correlations

We have so far established that the models we trained to estimate SES from aerial tiles derive their predictions mostly from features contained within residential areas. Even though the previous analysis sheds light on the features behind the models prediction for both low and high SES, it is still unclear how the importance of a residential area is affected by the fact it neighbors other types of urban classes. In order to address this question, we now focused on the spectrum of coactivations for both low ($d_{0,1}$) and high ($d_{3,4}$) SES. Mimicking the former analysis, we provide as a means of comparison the co-appearance gains $(\hat{g}_{0,1}^{(u,v)})_{(u,v)}$ and $(\hat{g}_{3,4}^{(u,v)})_{(u,v)}$ describing how much more/less likely a polygon of class u is to be within a low or high SES tile given that it co-occurs with a polygon of class v. As we focus on residential areas, we restrain u to be either a very low, low, medium, high or very high density residential area (isolated residential areas are left out since they are less likely to co-ocurr with other urban classes). Coactivations and co-appearance gains for the city of Paris are reported in Figures 5-7 while results for other cities may be found in the SI.

In terms of the models coactivations, we find surprisingly that models low SES activations of residential areas are the highest with respect to the reference, whenever these areas co-occur with clear non-residential ones. More surprisingly perhaps is the fact that these trends also take place when we consider high SES activations. What these results suggests is that whenever the model is presented with a sample containing residential areas next to non-residential ones, it tends to focus more on the former ones than whenever they occur by themselves or with other residential areas, indeed confirming that these types of models base their prediction mostly on elements contained within residential areas. This is perhaps most surprising when we consider that the type of neighbouring urban classes are quite determinant of the socioeconomic status of a given residential area. For instance, in the city of Paris, residential areas located next to infrastructures such as ports, railways or motorways are between 20 and 90% more likely to be of low SES than those either co-ocurring with other areas or that are by themselves. Even more, low density urban areas located next to very high density ones are between 90 and 264% more likely to be of low SES than in the reference, when considering all cities together. This further confirms that the model ignores differences among residential areas. Conversely, if we were to look at high SES coappearance gains, residential areas located next to nature related

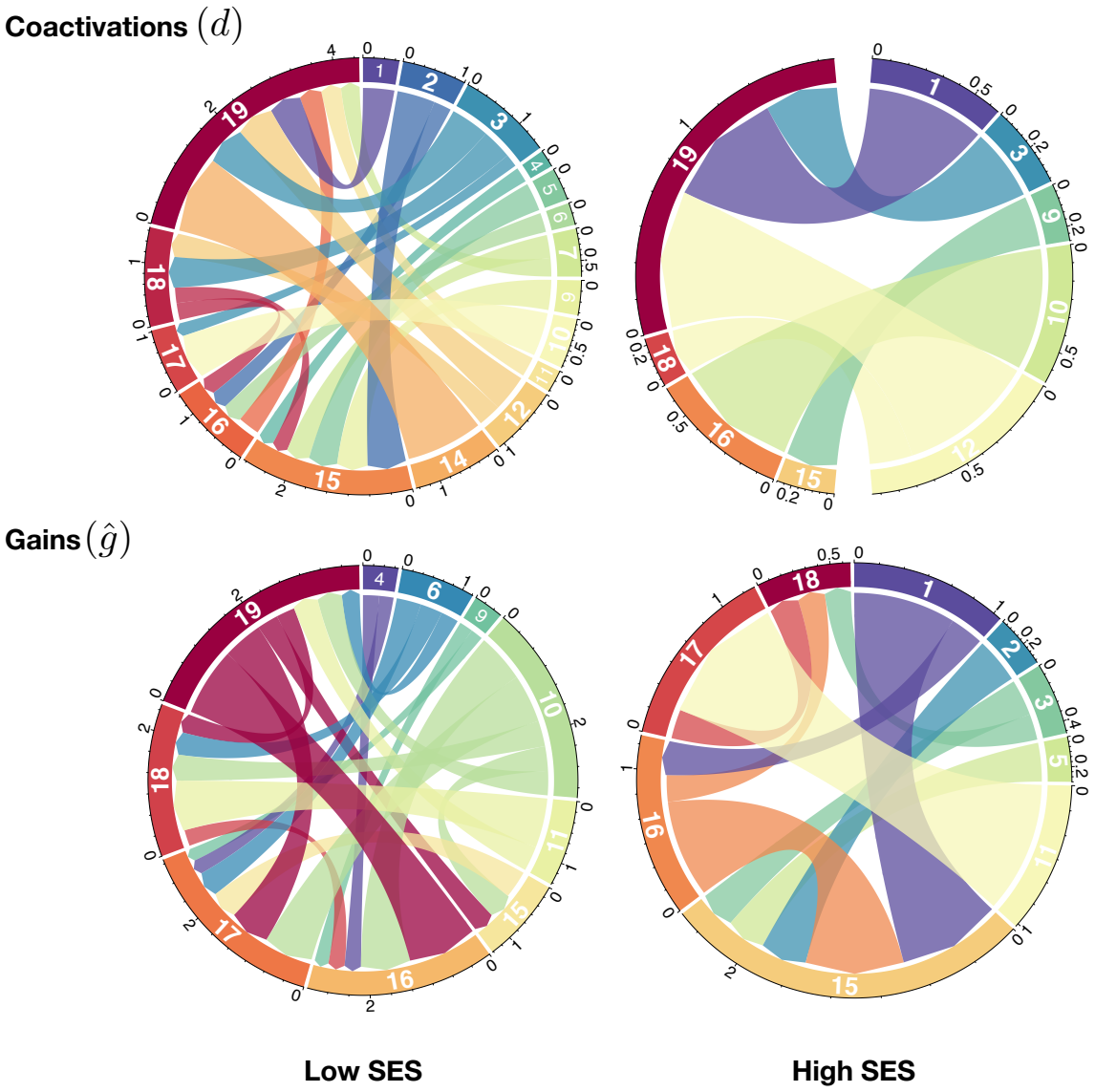


Figure 5-7: Chord diagrams of coactivations (top) and coappearance gains (bottom) estimated for both low (left) and high (right) SES in the city of Paris. Whenever urban class i appears more activated or is more likely to be of a given SES in the presence of urban class j than when appearing by itself or any other random urban class, it is represented in the diagram as $j \rightarrow i$. Each circular segment represents a urban class i connected by chords with width proportional to the coactivation/coappearance gain. Urban classes are labeled based on their ID (Table 5.1). Colors are assigned in order of appearance for better visualization based on a common colormap and are not exclusive of a given class. Also, to avoid overbearing results, only pairs of classes with values for both coactivations and coappearance gains of over 0.2 (at least 20% improvement over the univariate case) are shown.

scenery such as water, natural areas or agricultural land are between 20 and 110% more likely to be of high ses than in the reference, reinstating what is said above and invalidating **(H1)**.

5.5 Discussion

The successful deployment of urban policies aimed at curbing social problems as income inequality and poverty requires and updated and upscaled socioeconomic description of the city. To do so, policy makers are increasingly turning to deep learning based solutions despite their inherent lack of interpretability [191].

Our aim in this study has been to address this shortfall by 1) building a dataset from open data sources susceptible to be used for deep learning based urban solutions 2) providing a CNN-based framework for predicting socioeconomic status from aerial imagery using only open datasets for five French cities and 3) examining the activations derived from these models in terms of land cover. In doing so, we have built a model able to generate these predictions with state-of-the-art performance and have shown that models trained under these conditions seem to rely on features contained mostly within residential areas rather than non-residential ones. Finally, we have also observed that this pattern of activations differs significantly from the statistical information one can derive by joining land cover and socioeconomic information.
Our contribution opens up directions of research in the growing body of literature interested in the link between urban topology, architecture and socioeconomic status inference. It is not clear for instance how previously shown similarities between urban environments [5] affect the end prediction of socioeconomic status as well as the transferability of the learned models between cities. Further tests on different models, resolutions and interpretatibility techniques are necessary to reach the end-goal of deploying these models with a more complete understanding of their inner mechanisms.

Chapter 6

Conclusion

In this dissertation, we considered the task of detecting and developing tools making use of the socioeconomic correlations existing with respect to language, social networks and urban areas. The ever-expanding collection of social datasets that are now available hold an enormous potential to radically change the way in which we detect and update our societal knowledge on the formation and maintaining of socioeconomic inequalities. Leveraging the information contained within can nonetheless be a sisyphean task. Therefore, fulfilling the promise this sea of information holds calls for the development of well-designed and fine-tuned models able to accommodate the high-dimensional, multi-layered nature of social data and extract precise and reliable socioeconomic information. Through these series of works, we aimed at pushing current boundaries in that direction. To that end, we introduced a number of novel inference platforms that, fueled by the latent socioeconomic correlations that sway our behavior (Chapter 2), are able to provide an accurate socioeconomic portrait of society. In doing so, they rely on machine and deep learning algorithms that are able to process language (Chapter 3), network (Chapter 4) and neighborhood (Chapter 5) information and generate accurate predictions on individual attributes. Despite the progress that has ensued this work regarding the computational detection of socioeconomic inequalities many questions remain yet unanswered.

Disentangled network embedding to understand link formation

A logical next step regarding the network representation learning work here conducted is the understanding of disentanglement between network structure and node features. Indeed, in order to generate predictions on whether a given link between two users exist, it is still unclear whether users are linked together because they share similar features or rather they share similar features because they are linked. The framework introduced in Chapter 4 is able by design to take this information into account by modifying the information overlap between each. When applied to real-world social networks such as the one used in Chapter 2 aggregated with different

time-windows, our framework would be sensitive to changes in the amount of overlap needed to reach similar link prediction performances hence shedding light upon the homophily/social influence question.

Deep learning based detection of socioeconomic improvement

In Chapter 5, we developed an end-to-end pipeline able to generate fine-grained and accurate socioeconomic maps from census cells based on their corresponding satellite image. This technology is actually well-poised to generate continuous updates on the improvement/deterioration of a given neighborhood thus enabling the setup of policies to encourage/reverse the currents trend. Therefore, a key challenge associated to such techniques is to actually ensure that the model learned with a given snapshot of the country is still fit to predict socioeconomic inequalities later in time. In order to do so, we would then need to verify whether our model is able to generate reliable predictions at different timestamps or whether it needs to be adapted or retrained to ensure this capability. More generally, the application of deep learning technologies for socioeconomic inference can actually be plagued by issues related inherently to their design, particularly in terms of bias, interpretability and accessibility. In terms of the former two, most models learn associations between features originally contained in the dataset to give back predictions. In doing so, deep learning models are quite susceptible to silently learn and replicate biases originally contained in the datasets we train them on. This becomes even more problematic when we consider the task at hand and the uses that may be made from our predictions ,e.g , providing economical relief to socioeconomically deprived areas or deciding whether or not to open a business in a given neighborhood based on the predicted wealth of its inhabitants. Even if significant progress in this direction is currently taking place [232], more work needs to be done to address not only how to mitigate inherent biases but also on reacting on how directing aid to given areas might change the behaviour of its inhabitants and in turn the predictions our systems made. Regarding accessibility, currently, training large deep learning models requires significant amounts of computing power. This limits the type of work conducted here to few centers with enough resources to actually train our models. In order to further democratize their development and use, a greater amount of emphasis needs to be put in the design of smaller and faster models able to compete with state-of-the-art architectures such as the ones used in Chapter 5.

Transferable models to apply to data-sparse environments

Satellite to socioeconomic status computer vision models tend to function especially well in data rich environments and especially at fine-grained spatial resolutions. However, the settings were this technology would be most beneficial are actually lo-

cated in notorious data-scarce environments. Although much effort has already been invested in overcoming this key challenge, a direction worth exploring is the degree to which models trained in a given area might be able to be re-used in areas for which the model saw no data during training. In order to carry this to term, an interesting starting point would be to examine the spatial distributions of the error made by a model trained in a given country and tasked with predicting SES in similar regions. Once this task is completed, we would be able to understand the design features that need to be implemented to actually export these frameworks to the places they are needed the most.

Finally, I'm convinced that the potential that these models hold to improve our understanding and view of societal problems is great but not yet fulfilled. Granted, our ability to perform detailed and fine-grained analysis of society at the scale presented in this work, has never been so widespread. However, to truly develop models able to work within the efficiency and unbiased constraints we need them to, as well as to fully comprehend the hidden mechanisms that govern social processes, we need to further invest in diversifying the datasets we rely on and favor models not only in terms of performance score but also based on the resources they require to run. Moreover, we need to bridge the gap existing between industry and academia to enable public research to gain similar insights to the ones that seem exclusively available within industry, both in terms of computing power and data capabilities. Only by addressing these issues will we be able to fully pave the way for the techniques developed here to fully reach the people needing them the most.

Appendix A

Satellite to socioeconomic status inference results for other cities

A.1 Socioeconomic Predictions for other cities

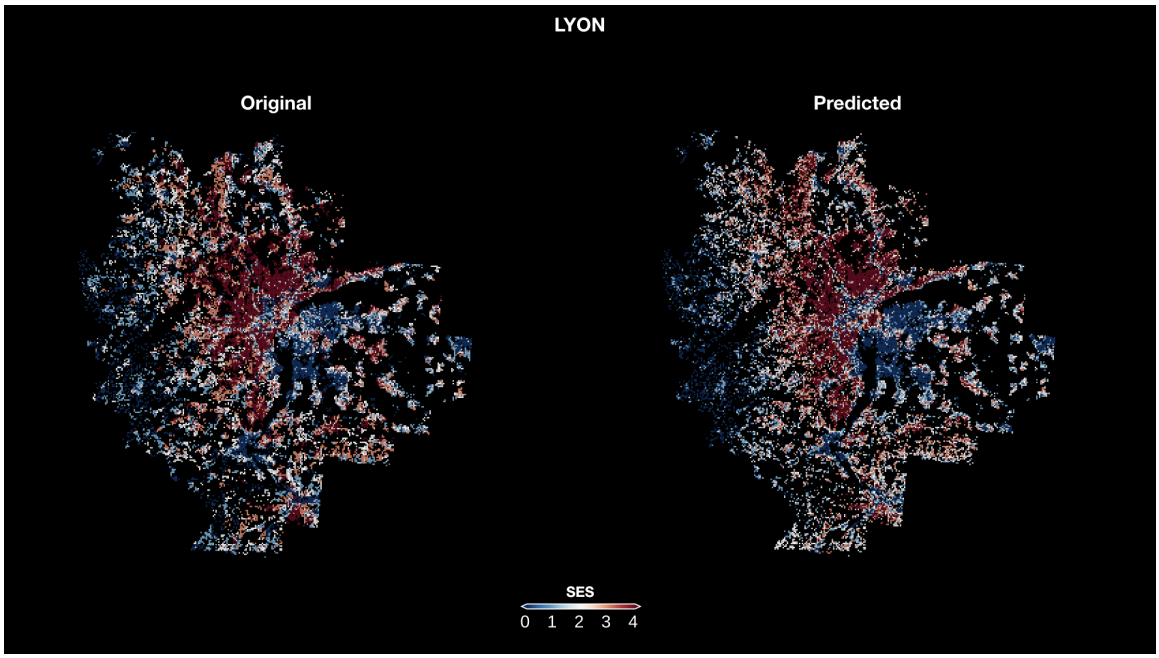


Figure A-1: Maps of observed and predicted average income for Lyon.

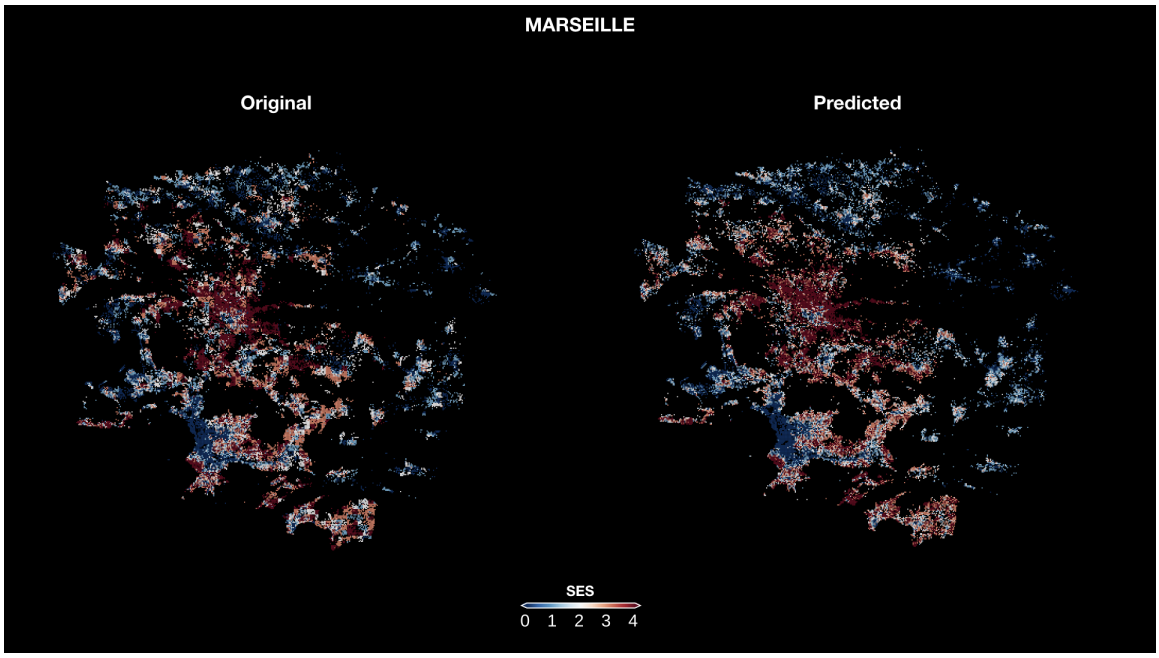


Figure A-2: Maps of observed and predicted average income for Marseille.

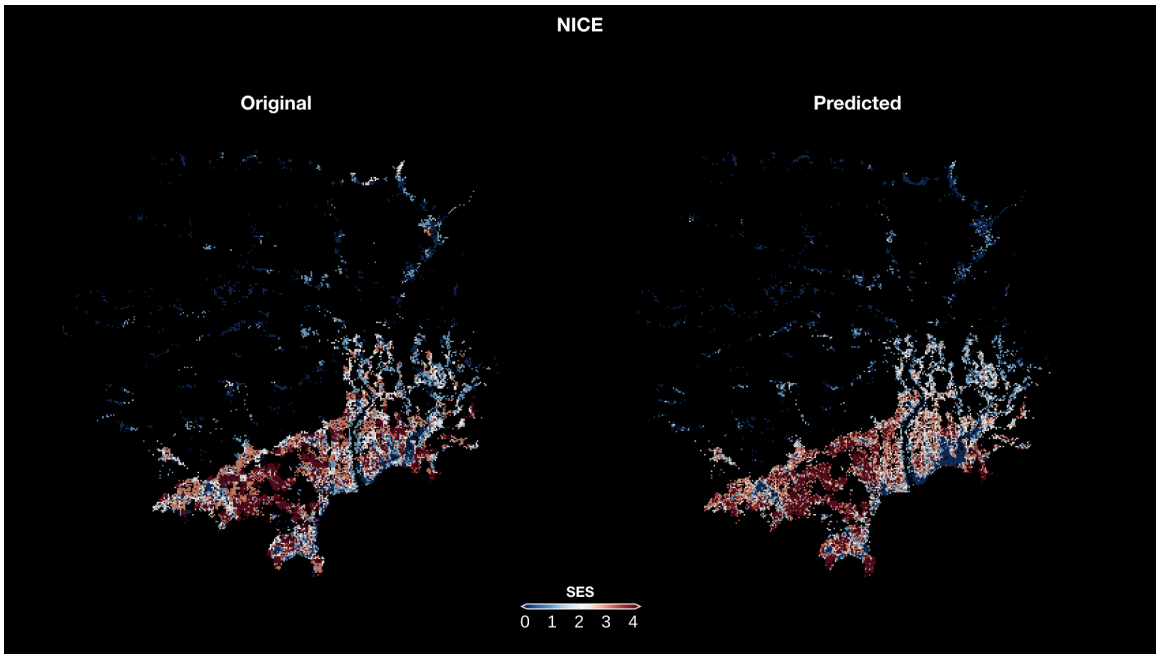


Figure A-3: Maps of observed and predicted average income for Nice.

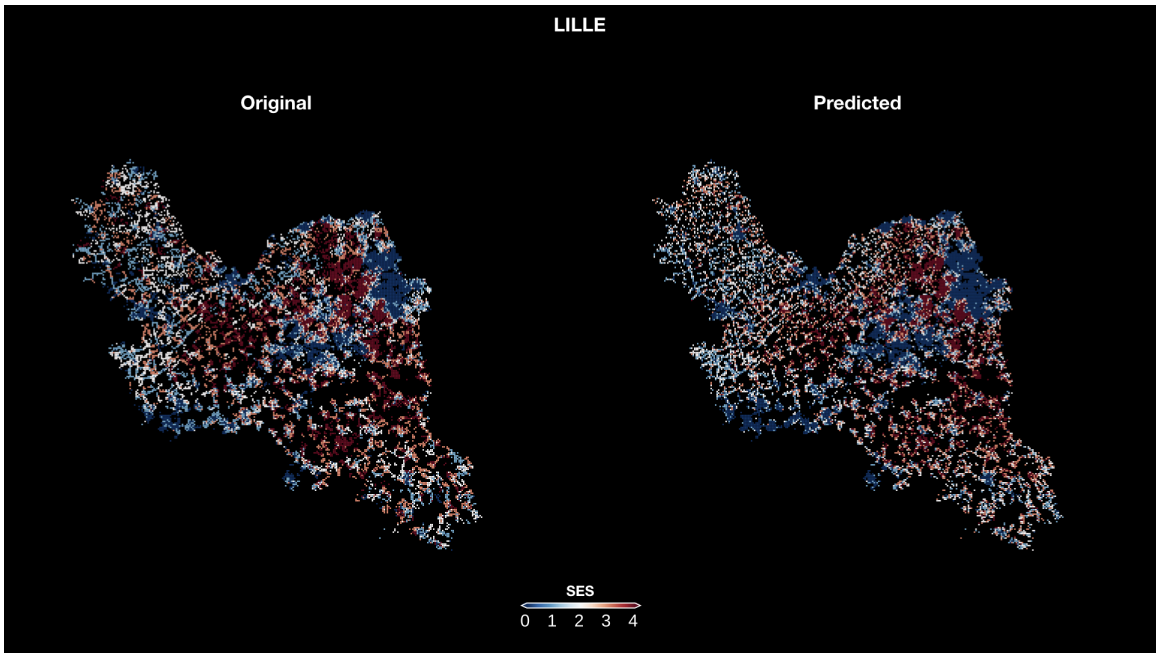


Figure A-4: Maps of observed and predicted average income for Lille.

A.2 Activation ratios & empirical probabilities

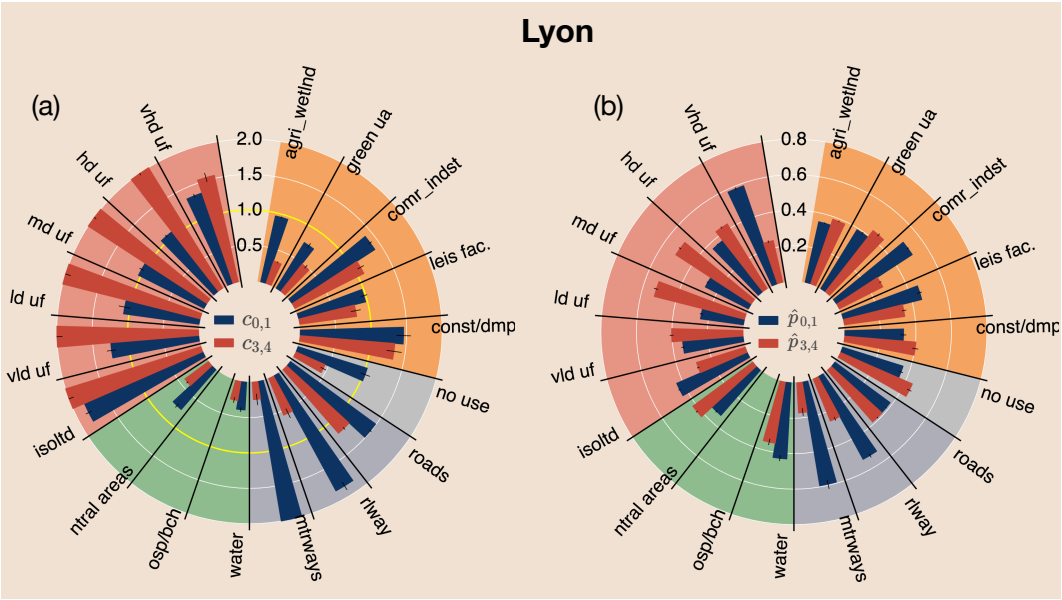


Figure A-5: Correlations between urban topology and socioeconomic status in the city of Lyon: (a) Mean model activation rate per urban class with 95% confidence interval for samples predicted as respectively as low SES (blue) or high SES (red) by the model (b) Estimated probability of a urban polygon belonging to the bottom or top quintile of the income distribution with bootstrapped 95% confidence interval.

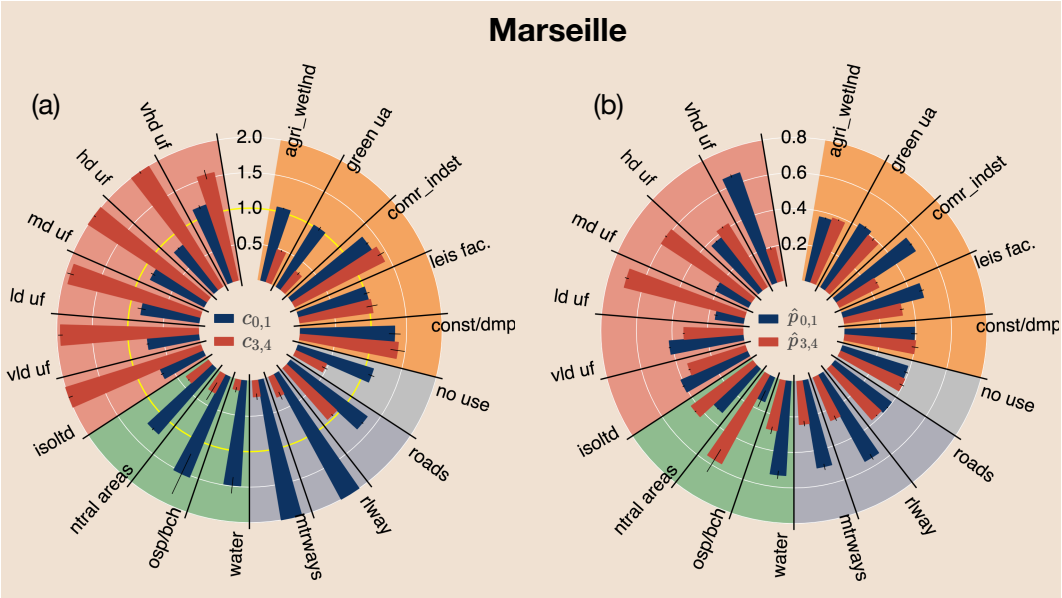


Figure A-6: Correlations between urban topology and socioeconomic status in the city of Marseille: (a) Mean model activation rate per urban class with 95% confidence interval for samples predicted as respectively as low SES (blue) or high SES (red) by the model (b) Estimated probability of a urban polygon belonging to the bottom or top quintile of the income distribution with bootstrapped 95% confidence interval.

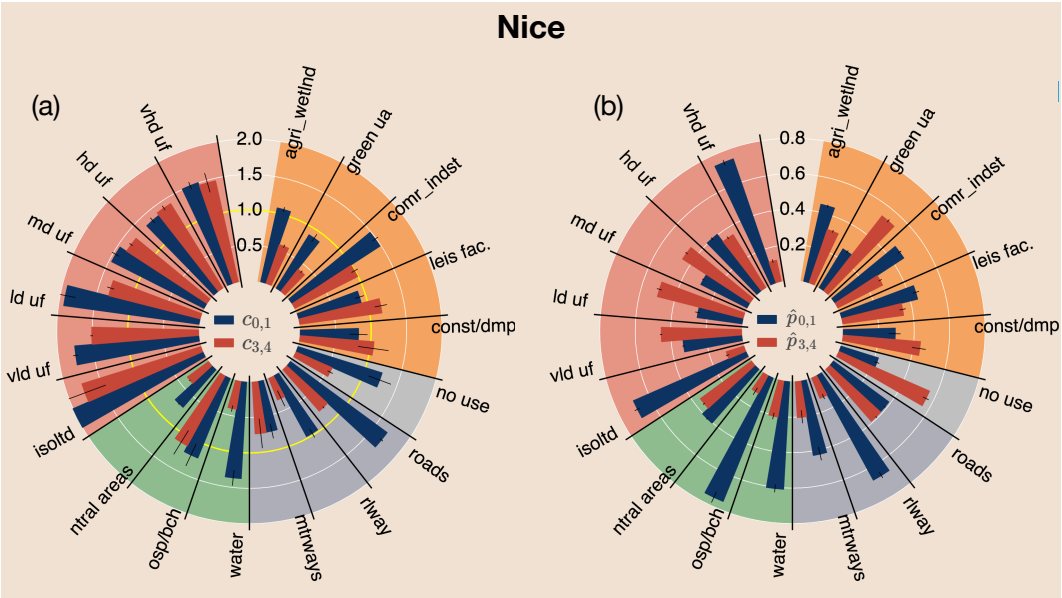


Figure A-7: Correlations between urban topology and socioeconomic status in the city of Nice: (a) Mean model activation rate per urban class with 95% confidence interval for samples predicted as respectively as low SES (blue) or high SES (red) by the model (b) Estimated probability of a urban polygon belonging to the bottom or top quintile of the income distribution with bootstrapped 95% confidence interval.

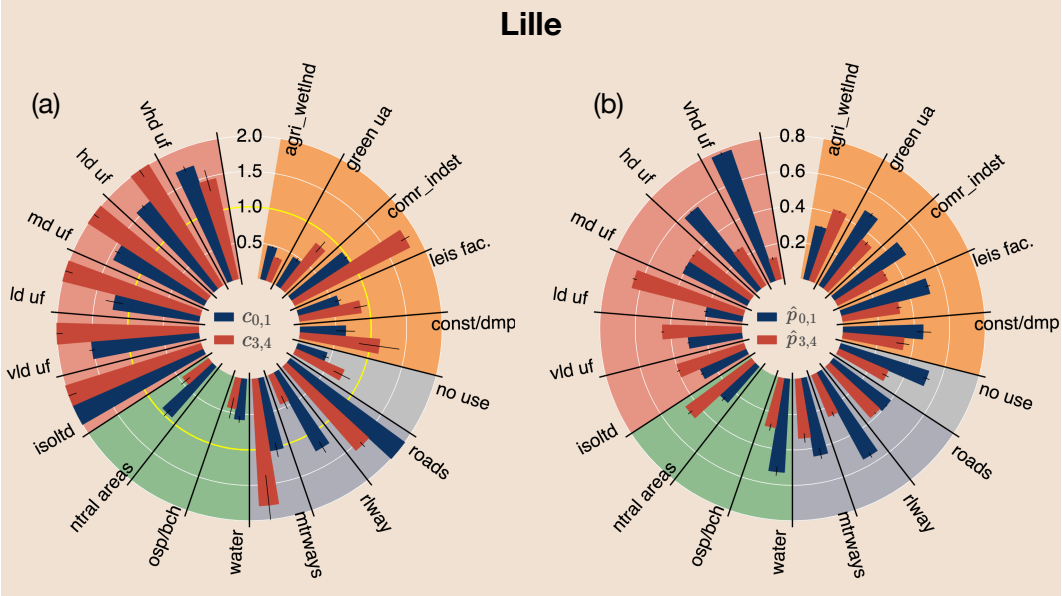


Figure A-8: Correlations between urban topology and socioeconomic status in the city of Lille: (a) Mean model activation rate per urban class with 95% confidence interval for samples predicted as respectively as low SES (blue) or high SES (red) by the model (b) Estimated probability of a urban polygon belonging to the bottom or top quintile of the income distribution with bootstrapped 95% confidence interval.

A.3 Co-Appearance Gains for other cities

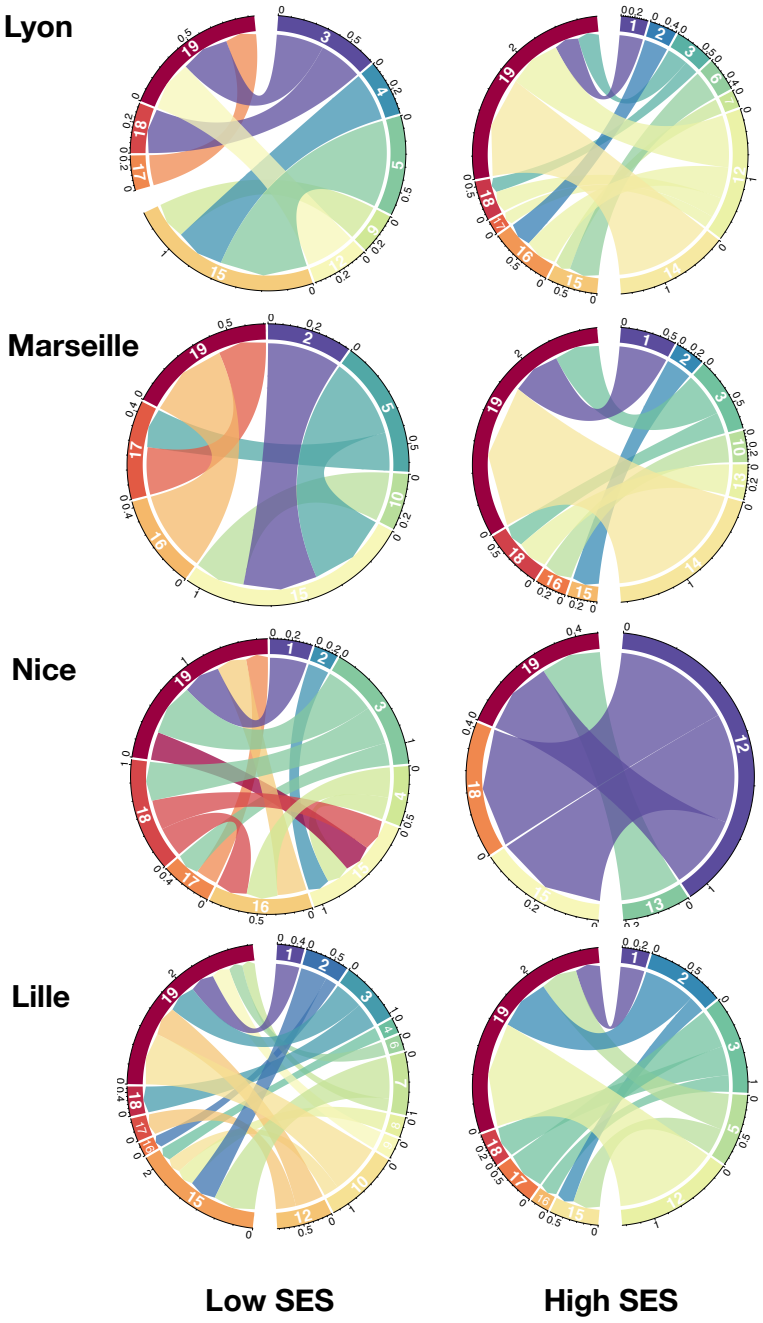


Figure A-9: Chord diagrams of coactivations estimated for both low (left) and high (right) SES in the remaining set of cities. Whenever urban class i appears more activated for a given SES in the presence of urban class j than when appearing by itself or any other random urban class, it is represented in the diagram as $j \rightarrow i$. Each circular segment represents a urban class i connected by chords with width proportional to the coactivation value. Urban classes are labeled based on their ID (Table 5.1). Colors are assigned in order of appearance for better visualization based on a common colormap and are not exclusive of a given class. Also, to avoid overbearing results, only pairs of classes with coactivations values of over 0.2 (at least 20% improvement over the univariate case) are shown.

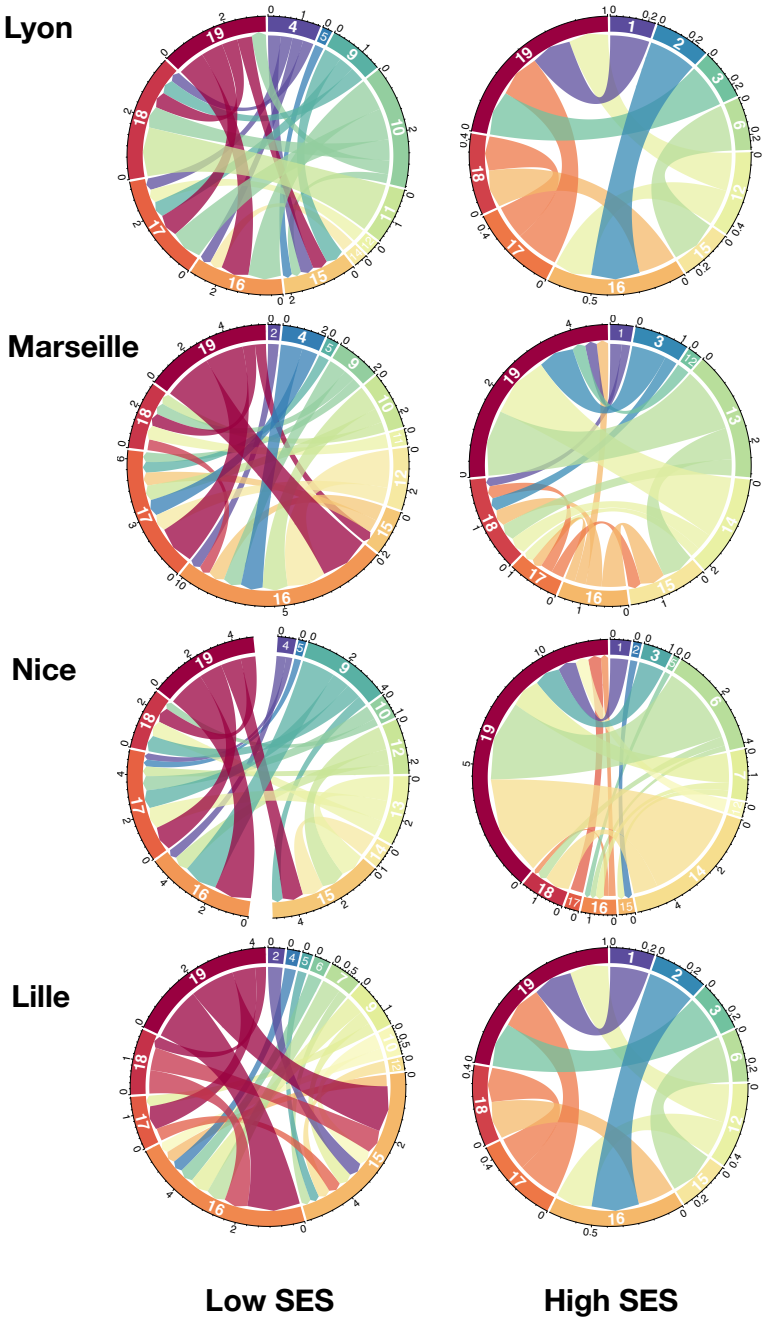


Figure A-10: Chord diagrams of coappearance gains estimated for both low (left) and high (right) SES in the remaining set of cities. Whenever urban class i is more likely to be of a given SES in the presence of urban class j than when appearing by itself or any other random urban class, it is represented in the diagram as $j \rightarrow i$. Each circular segment represents a urban class i connected by chords with width proportional to the coappearance gain. Urban classes are labeled based on their ID (Table 5.1). Colors are assigned in order of appearance for better visualization based on a common colormap and are not exclusive of a given class. Also, to avoid overbearing results, only pairs of classes with coappearance gains of over 0.2 (at least 20% improvement over the univariate case) are shown.

Bibliography

[1] Linkedinhelper, 2016. Accessed: 3 Nov. 2017. [Online]. Available at https://linkedhelper.com/.

[2] Jacob Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, and Eric Fleury. Socioeconomic dependencies of linguistic patterns in twitter: A multivariate analysis. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 1125–1134, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.

[3] Leona S. Aiken, Stephen G. West, and Raymond R. Reno. *Multiple regression: testing and interpreting interactions*. Sage Publications, 1991.

[4] Oluwaseun Ajao, Jun Hong, and Weiru Liu. A survey of location inference techniques on twitter. *J. Inf. Sci.*, 41(6):855–864, December 2015.

[5] Adrian Albert, Jasleen Kaur, and Marta C. González. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 1357–1366, 2017.

[6] Nikolaos Aletras and Benjamin Paul Chamberlain. Predicting twitter user socioeconomic attributes with network and language information. In *Proceedings of the 29th on Hypertext and Social Media*, HT '18, pages 20–24, New York, NY, USA, 2018. ACM.

[7] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[8] William J. Ashby. Un nouveau regard sur la chute du ne en français parlé tourangeau: s'agit-il d'un changement en cours? *Journal of French Language Studies*, 11(1):1–22, 2001.

[9] World Bank. Understanding poverty. https://www.worldbank.org/en/topic/poverty. Accessed: 2019-08-16.

[10] World Bank. Gini index, 2010. Accessed: 14 Jun. 2018. [Online]. Available at https://data.worldbank.org/indicator/SI.POV.GINI?locations=FR.

[11] Albert-László Barabási. *Bursts: The Hidden Pattern Behind Everything We Do*. Dutton Adult, 2010.

[12] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[13] Stefano Battiston, J. Doyne Farmer, Andreas Flache, Diego Garlaschelli, Andrew G. Haldane, Hans Heesterbeek, Cars Hommes, Carlo Jaeger, Robert May, and Marten Scheffer. Complexity theory and financial regulation. *Science*, 351(6275):818–819, 2016.

[14] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[15] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pages 153–160, 2007.

[16] Anita Berit Hansen and Isabelle Malderez. Le ne de négation en région parisienne : une étude en temps réel. *Langage et société*, 107(1):5–30, 2004.

[17] Basil Bernstein. Language and social class. *The British Journal of Sociology*, 11(3):271–276, 1960.

[18] Stephen van Bibber and Glenn Gilbert. Sociolinguistic theory: Linguistic variation and its social significance. *Language*, 72(3):617–620, 1996.

[19] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

[20] Joshua Evan Blumenstock. Fighting poverty with data. *Science*, 353(6301):753–754, 2016.

[21] Aleksandar Bojchevski and Stephan Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *International Conference on Learning Representations*, 2018.

[22] François Bourguignon. *The Globalization of Inequality*. Princeton University Press, 2015.

[23] Catherine Brissaud. *La réalisation de l'accord du participe passé employé avec avoir; de l'influence de quelques variables linguistiques et sociales*. Maison des sciences de l'homme, 1999.

[24] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[25] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18âĂŞ42, Jul 2017.

[26] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

[27] Jazmin L Brown-Iannuzzi, Kristjen B Lundberg, and Stephanie McKee. The politics of socioeconomic status: how socioeconomic status may influence political attitudes and engagement. *Current Opinion in Psychology*, 18:11 – 14, 2017. Inequality and social class.

[28] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral Networks and Locally Connected Networks on Graphs. In *arXiv:1312.6203 [cs]*, December 2013. arXiv: 1312.6203.

[29] Hongyun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616âĂŞ1637, Sep 2018.

[30] Kathryn Campbell-Kibler. New directions in sociolinguistic cognition. *University of Pennsylvania Working Papers in Linguistics*, 15(2):31–39, 12 2010.

[31] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 891–900, New York, NY, USA, 2015. ACM.

[32] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Deep neural networks for learning graph representations. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 1145–1152. AAAI Press, 2016.

[33] Marco Castelluccio, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. Training convolutional neural networks for semantic classification of remote sensing imagery. In *2017 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4, March 2017.

[34] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[35] Benjamin Paul Chamberlain, Clive Humby, and Marc Peter Deisenroth. Probabilistic inference of twitter users' age based on what they follow. *Lecture Notes in Computer Science*, pages 191–203, 2017.

[36] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[37] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 794–803, Stockholmsmässan, Stockholm Sweden, July 2018. PMLR.

[38] François Chollet et al. Keras. https://keras.io, 2015.

[39] Junyoung Chung, Caglar Gulcehre, Kyung Hyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[40] Gregory Clark. The surest cure for inequality. *The Wall Street Journal*, January 2017.

[41] Ryan Compton, David Jurgens, and David Allen. Geotagging one hundred million twitter accounts with total variation minimization. *2014 IEEE International Conference on Big Data (Big Data)*, Oct 2014.

[42] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. *International AAAI Conference on Web and Social Media*, 2011.

[43] Miles Corak. Income inequality, equality of opportunity, and intergenerational mobility. *Journal of Economic Perspectives*, 27(3):79–102, September 2013.

[44] Aron Culotta, Nirmal Kumar Ravi, and Jennifer Cutler. Predicting the demographics of twitter users from website traffic data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 72–78. AAAI Press, 2015.

[45] Marco De Nadai, Radu Laurentiu Vieriu, Gloria Zen, Stefan Dragicevic, Nikhil Naik, Michele Caraviello, César Augusto Hidalgo, Nicu Sebe, and Bruno Lepri. Are safer looking neighborhoods more lively? a multimodal investigation into urban life. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1127–1135. ACM, 2016.

[46] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016.

[47] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3844–3852, Red Hook, NY, USA, 2016. Curran Associates, Inc.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[49] Pascal Denis and Benoît Sagot. Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Lang. Resour. Eval.*, 46(4):721–736, December 2012.

[50] Lei Dong, Carlo Ratti, and Siqi Zheng. Predicting neighborhoods' socioeconomic attributes using restaurant data. *Proceedings of the National Academy of Sciences*, 2019.

[51] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

[52] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2224–2232, Red Hook, NY, USA, 2015. Curran Associates, Inc.

[53] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[54] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. Diffusion of lexical change in social media. *PLOS ONE*, 9(11):1–13, 11 2014.

[55] Martin Fixman, Ariel Berenstein, Jorge Brea, Martin Minnoni, and Carlos Sarraute. Inference of socioeconomic status in a communication graph. In *Argentine Symposium on Big Data (AGRANDA)*, pages 95–106, 09 2016.

[56] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[57] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.

[58] Harry G. Frankfurt. *On Inequality*. Princeton University Press, 2015.

[59] Neil Fraser. Sequence matcher python library, 2017. Accessed: 13 Dec. 2017.

[60] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data Clustering: Theory, Algorithms, and Applications (ASA-SIAM Series on Statistics and Applied Probability)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007.

[61] Hongchang Gao and Heng Huang. Deep Attributed Network Embedding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, IJCAI-18, pages 3364–3370, CA, USA, 2018. International Joint Conferences on Artificial Intelligence.

[62] Jian Gao, Yi-Cheng Zhang, and Tao Zhou. Computational socioeconomics. *Physics Reports*, 817:1 – 104, 2019. Computational Socioeconomics.

[63] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113, 2017.

[64] Jim Giles. Computational social science: Making the links. *Nature*, 488:448–50, 08 2012.

[65] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[66] Edward L. Glaeser and Abha Joshi-Ghani. *The urban imperative: towards competitive cities*. Oxford University Press, 2015.

[67] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, Sardinia, Italy, March 2010. PMLR.

[68] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

[69] Yoav Goldberg. A primer on neural network models for natural language processing. *J. Artif. Int. Res.*, 57(1):345–420, September 2016.

[70] Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[71] Google. Google maps static api, 2018. Accessed: 1 March. 2018. [Online]. Available at https://developers.google.com/maps/.

[72] Marc N. Gourevitch, Jessica K. Athens, Shoshanna E. Levine, Neil Kleiman, and Lorna E. Thorpe. City-level measures of health, health determinants, and equity to foster population health improvement: The city health dashboard. *American Journal of Public Health*, 109(4):585–592, 2019. PMID: 30789770.

[73] Mark Graham, Scott A. Hale, and Devin Gaffney. Where in the world are you? geolocation and language identification in twitter. *The Professional Geographer*, 66(4):568–578, 2014.

[74] Mark S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[75] Reilly N. Grant, David Kucher, Ana M. León, Jonathan F. Gemmell, Daniela S. Raicu, and Samah J. Fodeh. Automatic extraction of informal topics from online suicidal ideation. *BMC Bioinformatics*, 19(8):211, 2018.

[76] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

[77] John J Gumperz. The speech community. *Linguistic anthropology: A reader*, 1:66, 2009.

[78] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August 2016. Association for Computational Linguistics.

[79] William L. Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Eng. Bull.*, 40:52–74, 2017.

[80] William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017.

[81] Yuval N. Harari. *Sapiens : a brief history of humankind*. 1st U.S. edition. New York : Harper, 2015.

[82] Deborah Hardoon, Sophia Ayele, and Ricardo Fuentes-Nieva. An economy for the 1%: How privilege and power in the economy drive extreme inequality and how this can be stopped. *Oxford: Oxfam GB*, 2016.

[83] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[84] Oliver P. Hauser, Christian Hilbe, Krishnendu Chatterjee, and Martin A. Nowak. Social dilemmas among unequals. *Nature*, pages 1–4, 8 2019.

[85] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[86] Laurence Henry, Stephanie Barbu, Alban Lemasson, and Martine Hausberger. Dialects in animals: Evidence, development and potential functions. *Animal Behavior and Cognition*, 2, 05 2015.

[87] Philippe Hert. *Quasi-oralité de l'écriture électronique et sentiment de communauté dans les débats scientifiques en ligne*. Hermès Science Publications, 1999.

[88] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

[89] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[90] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[91] Erika Hoff. The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child development*, 74:1368–78, 10 2003.

[92] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[93] Ahmed Hosny, Chintan Parmar, Thibaud P. Coroller, Patrick Grossmann, Roman Zeleznik, Avnish Kumar, Johan Bussink, Robert J. Gillies, Raymond H. Mak, and Hugo J. W. L. Aerts. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLOS Medicine*, 15(11):1–25, 11 2018.

[94] Hadrien Hours, Eric Fleury, and Márton Karsai. Link prediction in the twitter mention network: Impacts of local structure and similarity of interest. In *ICDMW'16*, pages 454–461, 12 2016.

[95] Dirk Hovy, Anders Johannsen, and Anders Søgaard. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 452–461, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[96] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[97] Tianran Hu, Jiebo Luo, Henry Kautz, and Adam Sadilek. Home location inference from sparse and noisy data: Models and applications. In *Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, ICDMW '15, pages 1382–1387, Washington, DC, USA, 2015. IEEE Computer Society.

[98] Tianran Hu, Haoyuan Xiao, Jiebo Luo, and Thuy-vy Thi Nguyen. What the language you tweet says about your occupation. *International Conference on Web and Social Media*, 2016.

[99] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[100] Bernardo Huberman, Daniel Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), 2008.

[101] Janellen Huttenlocher, Marina Vasilyeva, Heidi R Waterfall, Jack L Vevea, and Larry V Hedges. *The varieties of speech to young children.*, volume 43. American Psychological Association, Huttenlocher, Janellen: Department of Psychology, University of Chicago, 5848 South University Avenue, Chicago, IL, US, 60637, hutt@uchicago.edu, 2007.

[102] Tomomichi Iizuka, Makoto Fukasawa, and Masashi Kameyama. Deep-learning-based imaging-classification identified cingulate island sign in dementia with lewy bodies. *Scientific Reports*, 9(1):8944, 2019.

[103] INSEE. Les salaires dans le secteur privé et les entreprises publiques., 2010. Accessed: 13 Dec. 2017. [Online]. Available at https://www.insee.fr/fr/statistiques/2122237/.

[104] INSEE. Données carroyées 200m https://www.insee.fr/fr/statistiques/2520034, 2016. Accessed: 2019-08-16.

[105] INSEE. Revenus, pauvreté et niveau de vie en 2014., 2017. Accessed: 1 Jul. 2017. [Online]. Available at https://www.insee.fr/fr/statistiques/3288151/.

[106] INSEE. Données carroyées, 2019. data retrieved from the Filosofi 2015 gridded data, https://www.insee.fr/fr/statistiques/4176290?sommaire=4176305.

[107] Asim Iqbal, Romesa Khan, and Theofanis Karayannis. Developing a brain atlas through deep learning. *Nature Machine Intelligence*, 1(6):277–287, 2019.

[108] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[109] Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. Understanding human mobility from twitter. *PLOS ONE*, 10(7):1–16, 07 2015.

[110] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651 – 666, 2010.

[111] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, December 2014. arXiv: 1412.6980.

[112] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[113] Katherine D. Kinzler, Emmanuel Dupoux, and Elizabeth S. Spelke. The native language of social cognition. *Proceedings of the National Academy of Sciences*, 104(30):12577–12580, 2007.

[114] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. *International Conference on Machine Learning*, pages 2693–2702, 2018.

[115] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *NIPS Workshop on Bayesian Deep Learning*, 2016.

[116] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[117] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[118] Gueorgi Kossinets and Duncan J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

[119] Gueorgi Kossinets and Duncan J. Watts. Origins of homophily in an evolving social network. *American Journal of Sociology*, 115:405–450, 2009.

[120] Jr. Kretzschmar, William A. Language Variation and Complex Systems. *American Speech*, 85(3):263–286, 08 2010.

[121] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[122] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 625–635, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[123] Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. Freshman or fresher? quantifying the geographic variation of language in online social media. *International Conference on Web and Social Media*, 2016.

[124] Vivek Kulkarni, Junting Ye, Steven Skiena, and William Yang Wang. Multi-view models for political ideology detection of news articles. *arXiv preprint arXiv:1809.03485*, 2018.

[125] Jussi M. Kumpula, Jukka-Pekka Onnela, Jari Saramäki, Kimmo Kaski, and János Kertész. Emergence of communities in weighted networks. *Phys. Rev. Lett.*, 99:228701, Nov 2007.

[126] Timothy La Fond and Jennifer Neville. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 601–610, New York, NY, USA, 2010. ACM.

[127] World Inequality Lab. World wealth and income database. https://wid.world/. Accessed: 2019-08-16.

[128] William. Labov. *Sociolinguistic patterns*. University of Pennsylvania Press Philadelphia, 1973.

[129] William Labov. *The Social Stratification of English in New York City*. Cambridge University Press, 2 edition, 2006.

[130] Bernard Laks. Why is there variation rather than nothing? *Language Sciences*, 39:31 – 53, 2013. Universalism and Variation in Phonology: Papers in Honour of Jacques Durand.

[131] Laurent, Guillaume, Saramäki, Jari, and Karsai, Márton. From calls to communities: a model for time-varying social networks. *Eur. Phys. J. B*, 88(11):301, 2015.

[132] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009.

[133] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436 EP –, 05 2015.

[134] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[135] Yannick Leo, Eric Fleury, J. Ignacio Alvarez-Hamelin, Carlos Sarraute, and Márton Karsai. Socioeconomic correlations and stratification in social-communication networks. *Journal of The Royal Society Interface*, 13(125):20160598, 2016.

[136] Yannick Leo, Márton Karsai, Carlos Sarraute, and Eric Fleury. Correlations of consumption patterns in social-economic networks. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '16, pages 493–500, Piscataway, NJ, USA, 2016. IEEE Press.

[137] Sébastien Lerique, Jacob Levy Abitbol, and Márton Karsai. Joint embedding of structure and features via graph convolutional networks, 2019.

[138] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2 edition, 2014.

[139] Michael KK Leung, Hui Yuan Xiong, Leo J Lee, and Brendan J Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics*, 30(12):i121–i129, 2014.

[140] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014.

[141] Jacob Levy Abitbol, Márton Karsai, and Eric Fleury. Location, occupation, and semantics based socioeconomic status inference on twitter. *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, Nov 2018.

[142] Jacob Levy Abitbol, Márton Karsai, and Eric Fleury. Optimal proxy selection for socioeconomic status inference on twitter. *Complexity*, 2019:1–15, 2019.

[143] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. Deepcas: An end-to-end predictor of information cascades. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 577–586, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[144] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, pages 6389–6399, 2018.

[145] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 556–559, 2003.

[146] Kent D. Van Liere and Riley E. Dunlap. The Social Bases of Environmental Concern: A Review of Hypotheses, Explanations and Empirical Evidence. *Public Opinion Quarterly*, 44(2):181–197, 01 1980.

[147] Justin Yifu Lin. *New Structural Economics : A Framework For Rethinking Development*. The World Bank, 2010.

[148] LinkedIn. Linkedin, 2018. Accessed: 31 May. 2018.

[149] Alejandro Llorente, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. Social media fingerprints of unemployment. *PLOS ONE*, 10(5):1–13, 05 2015.

[150] Linyuan Lu and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150 – 1170, 2011.

[151] Vincent Lucci and Agnès Millet. *L'orthographe de tous les jours. Enquête sur les pratiques orthographiques des Français*. ENS Editions, 1995.

[152] Shaojun Luo, Flaviano Morone, Carlos Sarraute, Matías Travizano, and Hernán A. Makse. Inferring personal economic status from social network location. *Nature Communications*, 8:15227 EP –, 05 2017.

[153] Pilar Manzanares-Lopez, Juan Pedro Muñoz-Gea, and Jose Maria Malgosa-Sanahuja. Analysis of linkedin privacy settings-are they sufficient, insufficient or just unknown? In *Proc. of the 10th International Conference on Web Information Systems and Technologies*, pages 285–293, 2014.

[154] Emile Mathieu, Tom Rainforth, N Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412, 2019.

[155] Viktor Mayer-Schnberger. *Big Data: A Revolution That Will Transform How We Live, Work and Think. Viktor Mayer-Schnberger and Kenneth Cukier*. John Murray Publishers, UK, 2013.

[156] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[157] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[158] Zhongqi Miao, Kaitlyn M. Gaynor, Jiayun Wang, Ziwei Liu, Oliver Muellerklein, Mohammad Sadegh Norouzzadeh, Alex McInturff, Rauri C. K. Bowie, Ran Nathan, Stella X. Yu, and Wayne M. Getz. Insights and approaches using deep learning to classify wildlife. *Scientific Reports*, 9(1):8137, 2019.

[159] Tomáš Mikolov, Anoop Deoras, Daniel Povey, Lukáš Burget, and Jan Černockỳ. Strategies for training large scale neural network language models. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 196–201. IEEE, 2011.

[160] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc.

[161] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[162] Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific Reports*, 3:1950 EP –, 06 2013.

[163] Boris Muzellec and Marco Cuturi. Generalizing point embeddings using the wasserstein space of elliptical distributions. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 10258–10269, USA, 2018. Curran Associates Inc.

[164] Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L. Glaeser, and César A. Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29):7571–7576, 2017.

[165] Vinod Nair and Geoffrey E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, Madison, WI, USA, 2010. Omnipress. event-place: Haifa, Israel.

[166] Galileo Namata, Ben London, Lise Getoor, and Bert Huang. Query-driven active surveying for collective classification. In *10th International Workshop on Mining and Learning with Graphs*, Edinburgh, Scotland, 2012.

[167] United Nations. Envision2030: 17 goals to transform the world for persons with disabilities. https://www.un.org/development/desa/disabilities/envision2030.html. Accessed: 2019-08-16.

[168] United Nations. 2018 revision of world urbanization prospects, 2018.

[169] Mark Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

[170] Vilfredo Pareto. *Manual of political economy*. Scholars Book Shelf, 1971.

[171] European Parliament. Eu statistics on income and living conditions methodology. https://ec.europa.eu/eurostat/statistics-explained/index.php/EU_statistics_on_income_and_living_conditions_(EU-SILC)_methodology. Accessed: 2019-08-16.

[172] Umashanthi Pavalanathan and Jacob Eisenstein. Confounds and consequences in geotagged twitter data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2138–2148, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[173] Steven Pearlstein. *Can American capitalism survive?: why greed is not good, opportunity is not equal, and fairness wont make us poor*. St. Martins Press, 2018.

[174] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, and Eduard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[175] Tiago P. Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Phys. Rev. X*, 4:011047, Mar 2014.

[176] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[177] Thomas Piketty and Arthur Goldhammer. *Capital in the Twenty-First Century*. Harvard University Press, 2014.

[178] A.P. Piontti, N. Perra, L. Rossi, N. Samay, A. Vespignani, C. Gioannini, M.F.C. Gomes, and B. Gonçalves. *Charting the Next Pandemic: Modeling Infectious Disease Spreading in the Data Science Age*. Springer International Publishing, 2018.

[179] JC Platt, D Simard, and PY Steinkraus. Best practices for convolutional neural networks. In *International Conference on Document Analysis and Recognition (ICDAR)*, page 958, 2003.

[180] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science advances*, 4(7):eaap7885, 2018.

[181] Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. Studying user income through language, behaviour and affect in social media. *PLOS ONE*, 10(9):1–17, 09 2015.

[182] Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. An analysis of the user occupational class through twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764, Beijing, China, July 2015. Association for Computational Linguistics.

[183] Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740, 2017.

[184] T. Douglas Price and Ofer Bar-Yosef. *Traces of Inequality at the Origins of Agriculture in the Ancient Near East*. Springer New York, New York, NY, 2010.

[185] Foster J. Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 445–453, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[186] Rajat Raina, Anand Madhavan, and Andrew Y Ng. Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, pages 873–880. ACM, 2009.

[187] Martin Ravallion. *The Economics of Poverty: History, Measurement, and Policy*. Oxford University Press, 2016.

[188] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[189] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1278–II–1286. JMLR.org, 2014.

[190] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

[191] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[192] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[193] Jari Saramäki, E. A. Leicht, Eduardo López, Sam G. B. Roberts, Felix Reed-Tsochas, and Robin I. M. Dunbar. Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, 111(3):942–947, 2014.

[194] P. Saunders. *Social Class and Stratification*. Society now. Routledge, 1990.

[195] Sanja Sćepanović, Igor Mishkovski, Bruno Gonçalves, Trung Hieu Nguyen, and Pan Hui. Semantic homophily in online communication: Evidence from twitter. *Online Social Networks and Media*, 2:1 – 18, 2017.

[196] Walter Scheidel. *The Great Leveler: Violence and the History of Inequality from the Stone Age to the Twenty-First Century*. Princeton University Press, 2017.

[197] Kenneth Scheve and David Stasavage. Democracy, war, and wealth: Lessons from two centuries of inheritance taxation. *American Political Science Review*, 106(1):81–102, 2012.

[198] Daniel A. Schult. Exploring network structure, dynamics, and function using networkx. In *In Proceedings of the 7th Python in Science Conference (SciPy*, pages 11–15, 2008.

[199] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8(9):1–16, 09 2013.

[200] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Oct 2017.

[201] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93, Sep. 2008.

[202] Copernicus Land Monitoring Service. Urban atlas, 2012. data retrieved from the France container, https://land.copernicus.eu/local/urban-atlas/urban-atlas-2012.

[203] Enya Shen, Zhidong Cao, Changqing Zou, and Jianmin Wang. Flexible Attributed Network Embedding. *arXiv:1811.10789 [cs]*, November 2018. arXiv: 1811.10789.

[204] Wesley Shrum, Neil H. Cheek, and Saundra MacD. Hunter. Friendship in school: Gender and racial homophily. *Sociology of Education*, 61(4):227–239, 1988.

[205] Doron Shultziner, Thomas Stevens, Martin Stevens, Brian Stewart, Rebecca Hannagan, and Giulia Saltini-Semerari. The causes and scope of political egalitarianism during the last glacial: A multi-disciplinary perspective. *Biology and Philosophy*, 25:319–346, 04 2010.

[206] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[207] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[208] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[209] John Steele Gordon. A short history of census taking. *The Wall Street Journal*, June 2019.

[210] Esra Suel, John W. Polak, James E. Bennett, and Majid Ezzati. Measuring social, environmental and health inequalities using deep learning and street imagery. *Scientific Reports*, 9(1):6229, 2019.

[211] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.

[212] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[213] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[214] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019.

[215] Lei Tang and Huan Liu. Leveraging social media networks for classification. *Data Mining and Knowledge Discovery*, 23(3):447–478, November 2011.

[216] Ziqi Tang, Kangway V. Chuang, Charles DeCarli, Lee-Way Jin, Laurel Beckett, Michael J. Keiser, and Brittany N. Dugger. Interpretable classification of alzheimer's disease pathologies with a convolutional neural network pipeline. *Nature Communications*, 10(1):2173, 2019.

[217] Techcrunch. Google revamps its google maps developer platform, 2018. https://techcrunch.com/2018/05/02/google-revamps-its-google-maps-developer-platform/.

[218] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[219] Yosuke Toda and Fumio Okura. How convolutional neural networks diagnose plant disease. *Plant Phenomics*, 2019, 03 2019.

[220] L. Torlay, Marcela Perrone-Bertolotti, Elizabeth Thomas, and Monica Baciu. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain Informatics*, pages 1–11, April 2017.

[221] Corinne Totereau, Catherine Brissaud, Caroline Reilhac, and Marie-Line Bosse. L'orthographe grammaticale au collège : une approche sociodifférenciée. *ANAE - Approche Neuropsychologique des Apprentissages chez l'Enfant*, 123:164–171, 09 2013.

[222] Phi Vu Tran. Learning to make predictions on graphs with autoencoders. In *5th IEEE International Conference on Data Science and Advanced Analytics*, 2018.

[223] Srinivas C Turaga, Joseph F Murray, Viren Jain, Fabian Roth, Moritz Helmstaedter, Kevin Briggman, Winfried Denk, and H Sebastian Seung. Convolutional networks can learn to generate affinity graphs for image segmentation. *Neural computation*, 22(2):511–538, 2010.

[224] Twitter. Twitter open api., 2018. Accessed: 1 March. 2018. [Online]. Available at https://developer.twitter.com/en/docs.html.

[225] Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 186–196, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[226] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[227] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1225–1234, New York, NY, USA, 2016. ACM.

[228] Martijn Wieling, John Nerbonne, and R. Harald Baayen. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLOS ONE*, 6(9):1–14, 09 2011.

[229] Yang Xu, Alexander Belyi, Iva Bojic, and Carlo Ratti. Human mobility and socioeconomic status: Analysis of singapore and boston. *Computers, Environment and Urban Systems*, 72:51 – 67, 2018.

[230] Yi Yang and Shawn Newsam. Uc merced land use dataset, 2017. Accessed: 1 Jun. 2018. [Online]. Available at http://weegee.vision.ucmerced.edu/datasets/landuse.html.

[231] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '18*. ACM Press, 2018.

[232] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.

[233] Dingyuan Zhu, Peng Cui, Daixin Wang, and Wenwu Zhu. Deep variational network embedding in wasserstein space. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, pages 2827–2836, New York, NY, USA, 2018. ACM.

[234] George Kingsley Zipf. *Human behavior and the principle of least effort.* Addison-Wesley Press, 1949.