Chaos ARTICLE scitation.org/journal/cha

Dynamics of fintech terms in news and blogs and specialization of companies of the fintech industry

Cite as: Chaos **30**, 000000 (2020); doi: 10.1063/5.0004487
Submitted: 12 February 2020 · Accepted: 4 July 2020 ·
Published Online: [redacted] 2020

View Online Export Citation CrossMark

Fabio Ciulla[1] and Rosario N. Mantegna[2,3,4,a)]

AFFILIATIONS

[1] Quid, San Francisco, California 94111, USA
[2] Dipartimento di Fisica e Chimica, Università di Palermo, 90128 Palermo, Italy
[3] Complexity Science Hub Vienna, 1080 Vienna, Austria
[4] Department of Computer Science, University College London, WC1E 6EA London, United Kingdom

Note: This article is part of the Focus Issue, Dynamics of Social Systems.
[a)] **Author to whom correspondence should be addressed:** rosario.mantegna@unipa.it

ABSTRACT

We perform a large scale analysis of a list of fintech terms in (i) news and blogs in the English language and (ii) professional descriptions of companies operating in many countries. The occurrence and the co-occurrence of fintech terms and locutions show a progressive evolution of the list of fintech terms in a compact and coherent set of terms used worldwide to describe fintech business activities. By using methods of complex networks that are specifically designed to deal with heterogeneous systems, our analysis of a large set of professional descriptions of companies shows that companies having fintech terms in their description present over-expressions of specific attributes of country, municipality, and economic sector. By using the approach of statistically validated networks, we detect geographical and economic over-expressions of a set of companies related to the multi-industry, geographically, and economically distributed fintech movement.

We present a study of the rapid development of a highly innovative industry. Specifically, we investigate the fintech industry, i.e., the industry developing technological innovations, technology-based products, and services for the financial sector. This industry presents a rather fast dynamics and a worldwide diffusion. These aspects make an analysis based on a big data approach very difficult due to the unavoidable variety, biases, and inconsistencies of the best available databases. In our study, we overcome these limitations by using the methodology of statistically validated networks (SVNs). In fact, this methodology is able to highlight over-expressed relationships between pairs of elements of bipartite networks obtained from heterogeneous sets. By investigating a list of terms used in a large corpus of news and blogs and in a large collection of professional descriptions of companies working worldwide, and by using the methodology of statistically validated networks, we detect over-expressions of some fintech terms in the descriptions of companies with specific attributes of geographical location and of economic activity.

I. INTRODUCTION

Fintech is a term used by several organizations and academics. The term describes research, activities, products, practices, and services bridging finance, information technology, software development, computer science, and sociology. As for many fruitful and deep concepts, the term meaning is not static, nor is it fully or uniquely defined,[1] and several attempts have been made to properly frame the concept[2] and its evolution over time.[3] The first written record of the "fintech" term is found in an academic paper by Bettinger.[4] At that time, the term was essentially unnoticed and it was independently reformulated in the early 1990s to describe a project initiated by Citigroup to facilitate technological cooperation efforts.[3] The global financial crisis of 2008 and the success of new players delivering financial services by means of technological innovations, particularly in Asia and in emerging countries, have triggered enormous interest toward fintech challenges and solutions.

Fintech is today a rapidly growing business area that is active at the interface of many industries all over the world. Tools and services of fintech companies affect (or have a potential to affect) many traditional and new areas of finance. The impact of fintech companies also extends well beyond the field of finance. Examples are products and services such as the ones associated with the use of the blockchain in the food supply chain or in the monitoring of infectious diseases.

In this contribution, we aim at answering two scientific questions. The first question asks whether some terms referring to products, services, and methods are jointly used to describe fintech activities in news and blogs in recent years. We answer this question by investigating a large corpus of texts of news and blog sources written in English collected over the Internet during the years from 2014 to 2018. The corpus is investigated with basic tools of network science.[5–7] Specifically, starting from a list of terms (composed of single or multiple words) highlighted by experts, we investigated the network of co-occurrence of pairs of terms in a large corpus of texts of news and blogs for each calendar year of the database. We verify that the network of co-occurrence becomes progressively more dense and topologically compact supporting the hypothesis that this group of terms describes business and technological activities addressed by the general term fintech.

The second scientific question focuses on the profile of companies with fintech interests or activities operating in many countries. Specifically, we investigate economic sector, country, and municipality of a very large number of companies located worldwide by using the list of terms selected in the first part of our study and by detecting their presence in the descriptions of companies that are present in the professional databases Capital IQ and Crunchbase. We show that the over-expression of economic sector, country (more precisely country or dependent territory), and municipality of the headquarter of the company presents two statistical regularities: (i) some companies dealing with fintech processes, products, or methods specialize on specific fintech sub-topics; (ii) some companies concentrate their activities in specific economic sectors and/or in specific geographical clusters.

This second investigation presents an important challenge due to the fact that the coverage of the databases is geographically heterogeneous with a special focus on western countries. To overcome this problem of bias of databases toward western countries, we leverage on a methodology developed in network science.[8,9] This methodology is based on the study of statistically validated networks,[9,10] and it is able to detect over-expressions of linkages in heterogeneous networks successfully overcoming the problem of the heterogeneity and bias of the coverage of databases.

By applying the methodology of statistically validated networks, we first construct three bipartite networks and we then analyze them to detect over-expressions of linkages that are present between (i) economic sectors, (ii) countries, and (iii) municipalities of companies and fintech terms characterizing different areas of fintech products, services, and activities such as, for example, *financial inclusion*, *anti-money laundering* (AML), etc. In other words, our methodology highlights specializations of sets of companies in an heterogenous setting, allowing us to obtain statistically significant results starting from a heterogeneous source of data.

The paper is organized as follows. In Sec. II, we describe a set of selected fintech terms and the investigated databases. Section III presents the empirical results obtained in the analysis of networks of co-occurrence of fintech terms sampled at different calendar years. In Sec. IV, we investigate over-expressions detected in the bipartite networks of (i) economic sectors and fintech terms, (ii) countries and fintech terms, and (iii) municipalities and fintech terms. Section V discusses the results obtained and presents some conclusions.

II. FINTECH TERMS AND DATASETS

In this paper, we investigate the occurrence and co-occurrence of a set of 53 fintech terms. The set is selected starting from the analysis of a series of fintech terms collected and commented by experts in several web pages. One example of these lists of terms can be accessed at the web page reporting the article "Fintech lingo explained" by Irrera and Caspani, https://www.reuters.com/article/us-usa-fintech-explainer-idUSKBN19D29I.[11] Other examples of web pages with fintech list of terms are (i) https://eba.europa.eu/financial-innovation-and-fintech/glossary-for-financial-innovation, (ii) https://www.nbs.sk/en/financial-market-supervision1/fintech/fintech-glossary, and (iii) https://www2.deloitte.com/uk/en/pages/financial-services/articles/fintech-glossary.html.

The 53 investigated terms are listed in Table I. They include (a) words like bitcoin, blockchain, and crowdfunding, (b) groups of words expressing a precise concept such as anti-money laundering, combating the financing of terrorism, etc., (c) word contractions such as fintech, finserv, and segwit (together with their expanded terms), and (d) acronyms [software as a service (SAAS) and Europay, MasterCard, and Visa (EMV)]. It is worth stressing that we have used acronyms only in the absence of polysemy. For example, we did not use the widely used acronym AML for anti-money laundering because it is also frequently used for acute myeloid leukemia, which is a distinct concept.

Our first investigation concerns the occurrence and co-occurrence of fintech terms in texts of a corpus of news and blogs. The database of news and blogs covers texts distributed over the Internet during the calendar years of 2014, 2015, 2016, 2017, and 2018. It consists of approximately 1×10^9 texts written in the English language collected by considering approximately 60 000 news sources and 500 000 blogs. The corpus is a proprietary corpus of the company LexisNexis. The geographical origin of text sources is primarily located in the United States (47.5% of texts) and in the United Kingdom (15.4% of texts). The remaining 37.1% of texts originates from 207 different sovereign countries or overseas territories or dependent territories or unincorporated territories such as, for example, Hong Kong, Macau, Greenland, Puerto Rico, Faroe islands, Falkland islands, etc. For the sake of simplicity, in Secs. III–V, we use the word country to describe an entity being a sovereign country or an overseas territory or a dependent territory or an unincorporated territory or a similar type of institution. In this corpus, we investigate the occurrence and co-occurrence of fintech terms to track the evolution of the use of our selected terms of fintech products and services in the English language in recent years.

In our second investigation, the occurrence of selected fintech terms is investigated in the professional description of companies operating in many countries. The dataset of company descriptions is a dataset curated by the Quid company. The dataset was obtained

TABLE I. List of fintech terms investigated in our study. Terms are listed in alphabetical order from the first to the third column. The terms in parenthesis are expanded variants of the previous term.

Anti-money laundering	Genesis block	Robo-advisors
Bitcoin	Hard fork	(automate investment advice)
Blockchain	Hash rate	SAAS
Card not present	High speed networks	(software as a service)
Chief data officer	Initial coin offering	Segwit
Collaborative consumption	Insurtech	(segregated witness)
Collaborative economy	Know your customer	Sharding
Combating the financing of terrorism	Knowledge-based authentication	Single sign-on authentication
Counter-terrorist financing	Messaging commerce	Smart contracts
Crowdfunding	on-Boarding	(blockchain-based contracts)
Cryptocurrency	Open banking	Social lending
Digital wallet	P2P lending	Soft fork
Distributed ledger technology	(peer-to-peer lending)	Sybil attack
EMV chip	Payment gateway	Token sale
(Europay, MasterCard, and Visa)	PCI compliance	Tokenization
Equity-crowdfunding	(payment card industry compliance)	Unbanked
Ethereum blockchain	Point-of-sale	Underbanked
Financial inclusion	Proof-of-authority	User as owner
Finserv	Proof-of-stake	Virtual currency
(financial services industry)	Proof-of-work	
Fintech	Regtech	
(financial technology)	(regulatory technology)	

by merging the information present in two proprietary databases. These databases are the Capital IQ database of S&P Global company and the Crunchbase Pro database of Crunchbase company. Capital IQ database provides a quite complete coverage of publicly listed companies. In fact, the database covers 99% of global market capitalization according to Capital IQ website. Crunchbase database is more focused on innovative companies although currently also covers public and private companies on a global scale. Our dataset is obtained from the merging and pre-processing of the two databases. The total number of company descriptions is about 2.2×10^6. They are descriptions of companies with headquarters located in 239 different countries (where country has the broadly defined meaning clarified above) and classified as working in 68 different economic sectors. Although the dataset covers a large part of global market capitalization, it is not unbiased. In fact, a very high percent of companies are located in the United States (61.3%) and in the United Kingdom (7.50%) indicating that most small and innovative companies included in the datasets are operating in these two countries. Other top represented countries are China (2.48%), Germany (1.99%), France (1.76%), India (1.60%), Canada (1.51%), Italy (1.38%), Spain (1.35%), and Australia (1.28%). The bias is reduced but still present when we only consider public companies. For public companies, the ten top countries with highest percent of companies are United States (29.3%), Canada (10.3%), China (7.36%), India (6.32%), Japan (5.50%), United Kingdom (3.72%), Australia (3.51%), South Korea (3.25%), Taiwan (2.59%), and Hong Kong (2.37%). In our analysis, we therefore need to take into account the bias that is present in the dataset. In Sec. IV, we will take into account the bias by using a statistical methodology of network science that is able to highlight over-expression in bipartite networks in the presence of a pronounced heterogeneity of the elements (in the present case the attributes of companies). Both texts of news and blogs, as well as texts of companies' descriptions, have been indexed and queried using the open-core Elasticsearch search engine.

III. RESULTS ON THE ANALYSIS OF TEXTS OF NEWS AND BLOGS

We first search the fintech terms in the texts of the corpus of news and blogs for the calendar years from 2014 to 2018. The counts obtained are shown in Table II. The table shows that the occurrence of the 53 fintech terms is quite heterogeneous ranging from the 1 671 363 occurrences of *cryptocurrency* in 2018 to no occurrence of *user as owner* in 2017 and 2018. The pronounced heterogeneity is not too surprising due to the fact that the fintech list of terms comprises both quite wide concepts such as, for example, *software as a service* and very specialized concepts such as, for example, *hard fork* or *soft fork*. The number of texts investigated changes only moderately over the years. Their values are reported in the last row of Table II. The minimum number of texts investigated in a year was about 136×10^6 in 2014 and the maximum was about 183×10^6 in 2016. The average value was 167×10^6 with a standard deviation of 18.2×10^6, i.e., only about 11% of the average value. In the bottom part of Table II, we also show the total occurrence of fintech terms per year and the number of texts with at least one fintech term.

For some terms, we note a quite pronounced variation of the occurrence. For example, bitcoin, cryptocurrency, blockchain, smart contracts, insurtech, and regtech show prominent large variations of the occurrences in a relatively limited period of time. The occurrence analysis is, therefore, highlighting heterogeneity of the fintech

TABLE II. Occurrence of fintech terms in the texts of corpus of news and blogs for each investigated calendar year (from second to sixth column). Occurrence is ranked from top to bottom with the rank of the term defined by the total number of occurrences observed in the 5 years (seventh column). The last column, labeled as "Companies descriptions," shows the occurrence of the fintech terms in the professional documents describing companies included into Capital IQ and Crunchbase Pro databases.

Fintech term	News and blogs 2014	News and blogs 2015	News and blogs 2016	News and blogs 2017	News and blogs 2018	News and blogs All years	Companies descriptions
Software as a service (SAAS)	669 549	745 176	559 525	482 543	509 814	2 966 607	14 210
Bitcoin	196 728	158 893	127 020	385 084	1 595 799	2 463 524	1 785
Cryptocurrency	31 182	33 573	31 566	207 403	1 671 363	1 975 087	1 908
Blockchain	11 391	46 935	118 145	371 307	1 009 427	1 557 205	6 378
Fintech	89 435	197 436	321 873	421 670	498 991	1 529 405	5 331
Crowdfunding	201 681	288 203	253 103	223 953	222 131	1 189 071	1 996
Point-of-sale	267 858	275 910	209 134	186 231	203 124	1 142 257	5 230
Finserv	187 031	224 312	195 813	180 241	154 649	942 046	1 224
Anti-money laundering	46 586	60 800	73 999	76 564	96 464	354 413	359
Financial inclusion	38 048	54 993	69 253	73 368	86 089	321 751	268
Virtual currency	52 121	31 796	29 339	54 565	70 715	238 536	246
On-boarding	35 901	44 238	40 336	38 952	35 782	195 209	459
Proof-of-work	1 152	1 889	1 893	4 235	180 364	189 533	32
Smart contracts	523	3 221	12 160	39 983	105 688	161 575	521
Unbanked	27 147	30 378	29 342	32 052	39 973	158 892	222
Payment gateway	30 805	36 781	40 530	20 857	26 558	155 531	765
Digital wallet	29 101	22 795	21 976	24 242	30 001	128 115	194
Tokenization	21 083	34 056	18 966	20 855	29 927	124 887	173
Know your customer	15 455	18 448	19 941	24 062	34 547	112 453	135
P2P lending	15 812	24 963	30 043	18 377	19 765	108 960	382
Proof-of-stake	828	1 078	1 465	3 656	97 793	104 820	34
EMV chip	18 534	31 545	22 306	10 731	10 650	93 766	39
PCI compliance	25 918	27 098	11 582	8 542	8 129	81 269	194
Distributed ledger technology	20	2 122	10 954	22 064	44 991	80 151	147
Initial coin offering	256	3	1 100	23 440	46 168	70 967	63
Equity-crowdfunding	9 907	19 297	16 938	14 062	9 771	69 975	201
Insurtech	19	31	6 071	30 857	31v145	68 123	269
Ethereum blockchain	7	362	2 701	16 925	46 340	66 335	168
Underbanked	10 165	11 953	11 749	10 525	18 639	63 031	109
Token sale	8	212	79	23 079	32 848	56 226	47
Card not present	13 944	15 682	10 721	5 844	6 079	52 270	87
Robo-advisors	2 719	7 253	18 315	10 885	8 299	47 471	21
Regtech	1 455	4 153	6 233	16 116	19 139	47 096	137
Chief data officer	4 339	9 167	9 038	8 217	11 470	42 231	2
Open banking	282	671	2 733	11 227	23 122	38 035	47
High speed networks	5 547	6 328	4 233	4 403	4 227	24 738	37
Hard fork	22	148	709	6 013	17 161	24 053	2
Collaborative economy	2 125	4 575	2 914	1 851	1 537	13 002	47
Collaborative consumption	4 935	3 694	1 978	820	721	12 148	83
Sharding	2 949	2 301	1 258	1 631	3 823	11 962	17
Counter-terrorist financing	946	1 070	2 498	2 542	3 170	10 226	9
Segwit	5	22	260	3 825	5 129	9 241	2
Hash rate	896	461	275	1 201	5 605	8 438	4
Combating the financing of terrorism	1 072	790	1 756	2 012	2 518	8 148	2
Knowledge-based authentication	1 828	726	1 089	976	1 402	6 021	11
Single sign-on authentication	1 694	1 593	669	770	461	5 187	6
Genesis block	381	55	309	495	3 938	5 178	4
Social lending	608	599	829	1 011	720	3 767	31

TABLE II. *(continued.)*

Fintech term	News and blogs 2014	News and blogs 2015	News and blogs 2016	News and blogs 2017	News and blogs 2018	News and blogs All years	Companies descriptions
Proof-of-authority	30	55	193	272	1 407	1 957	2
Soft fork	7	10	210	688	910	1 825	1
Messaging commerce	26	9	43	327	49	454	0
Sybil attack	107	18	45	24	197	391	0
User as owner	1	3	1	0	0	5	0
Total occurrences of fintech terms	2 080 169	2 487 880	2 355 211	3 131 576	7 088 729		43 641
Texts with at least one fintech term	1 418 726	1 690 290	1 589 152	1 887 749	3 742 348		38 648
Number of texts in the corpus	136 048 047	172 912 445	182 959 692	169 448 198	175 559 955		2.2×10^6

terms and also a pronounced dynamics of some of them. We interpret this dynamics as an indication of the process of definition and specialization of the new terms. Let us consider, for example, the two terms fintech and finserv. These two terms are connoting different aspects of technological applications and service solutions of specific financial problems. The semantic difference between the two terms is debated over the years (see, for example, the 2015 blog https://finiculture.com/finserv-fintech/ for an opinion about it). The occurrence dynamics of the two terms observed from 2014 to 2018 shows a clear pattern. The term finserv has a pattern of decreasing occurrence while the reverse is true for fintech. In other words although in past few years, the two terms have been both used with a similar level of diffusion; in most recent years, fintech is emerging as the term describing both technological solutions and digital services applied to financial innovations.

The second type of investigation concerns the co-occurrence of pairs of fintech terms in the same text. In this investigation, we start to make use of networks as an analysis tool, indeed fintech terms are represented as nodes and an edge exists between two nodes when the two fintech terms are present in the same text at least once. In Table III, we show the time evolution of the number of nodes and edges of the network of co-occurrence of fintech terms. The table shows that the co-occurrence network is always characterized by a number of nodes very close to the number of investigated terms and by a number of edges that is growing from 2014 to 2018. In all years, we detect a single connected component and the network edge density is growing from 0.467 (in 2014) to 0.756 (in 2018). In parallel with the edge density increases, we also detect a steadily decrease of the average path length. The diameter of the network, i.e., the longest distance between any two terms in number of steps, is 3 for the 2014–2016 years and jumps to 2 in the last two years. The network is, therefore, highly dense and compact in the investigated years.

By performing numerical simulations, we have verified that the topology of the unweighted co-occurrence network is consistent with the one of an Erdös–Rényi model[6,7] with the same number of nodes and edges. However, the consistency of the empirical topology with an Erdös–Rényi topology does not mean that the co-occurrence of words is a random phenomenon. In fact, hereafter, we show that a null hypothesis of random matching of two different terms in the same text is not consistent with the observed value $N_{A,B}$ of co-occurrence of terms A and B. In our null model, the probability of occurrence of each term A is $P(A)$. By assuming a completely random matching of two terms A and B in the same text, the probability of observing a co-occurrence is the product of $P(A)$ times $P(B)$. Starting from this probability and assuming as a null model a binomial distribution with probability $P(A)P(B)$, the expected value $E[N_{A,B}]$ of the co-occurrence is given by $N_T P(A)P(B)$, where N_T is the total number of texts analyzed. The standard deviation of the same variable is $\sqrt{N_T P(A)P(B)(1 - P(A)P(B))}$. Under this null hypothesis, for each pair of terms, we estimate a z-score by computing

$$z(A,B) = \frac{N_{A,B} - E[N_{A,B}]}{SD[N_{A,B}]} = \frac{N_{A,B} - N_T P(A)P(B)}{\sqrt{N_T P(A)P(B)(1 - P(A)P(B))}}. \quad (1)$$

By analyzing the z-score values for all pairs of terms of the co-occurrence networks, we verify that z values are very large and in all cases, they exceed 3 for a fraction of edges ranging from 80.0% (in 2014) to 91.3% (in 2017). In summary, almost all detected co-occurrences of pairs of terms are not consistent with a random

TABLE III. Number of nodes, number of edges, edge density, number of connected components, average path length, and diameter of fintech term co-occurrence networks for each calendar year. The co-occurrence network of fintech terms is obtained by analyzing texts of a corpus of approximately 1×10^9 texts collected from news and blogs.

Year	No. nodes	No. edges	No. edge density	No. connected components	Average path length	Diameter
2014	46	483	0.467	1	1.54	3
2015	51	625	0.490	1	1.52	3
2016	52	823	0.621	1	1.38	3
2017	53	950	0.689	1	1.31	2
2018	52	1002	0.756	1	1.24	2

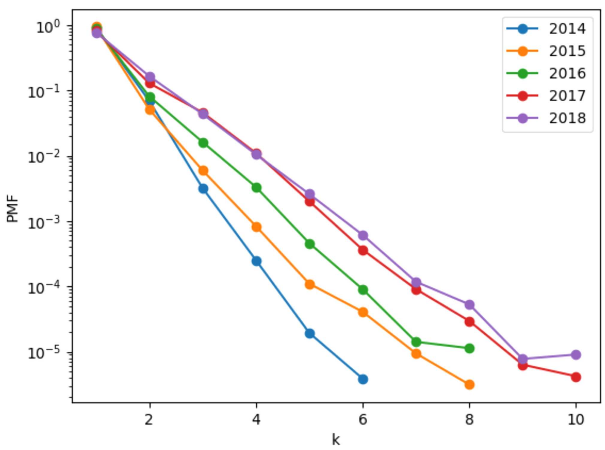

FIG. 1. Probability mass function of the number of co-occurrences of fintech terms detected in a single text. Symbols of different colors refer to different years. The vertical axis is logarithmic.

matching of the terms and they suggest that their joint use carry information in the text.

We also verify that the detected co-occurrences are not originated in a limited number of texts including the presence of many of the terms investigated. In Fig. 1, we show the probability mass function of observing k co-occurrences of fintech terms in a single text. The probability mass function is shown in a semi-logarithmic plot and it is well approximated by an exponentially decaying function. The figure shows that multiple co-occurrences increases in texts from 2014 to 2018 but the largest majority of texts presents just a single co-occurrence of fintech terms.

To further verify the role of the heterogeneity of the number of co-occurrences, we characterize the co-occurrence network as a weighted network where the weight of a link between node A and node B is given by the co-occurrence $N_{A,B}$. In this weighted network, we perform a community detection analysis with the algorithm Infomap[12] to search for any internal structure of the co-occurrence networks. The Infomap algorithm is one of the most widely used community detection algorithms. It can be applied both to unweighted and weighted networks. We apply the Infomap algorithm to the weighted co-occurrence networks and we find the

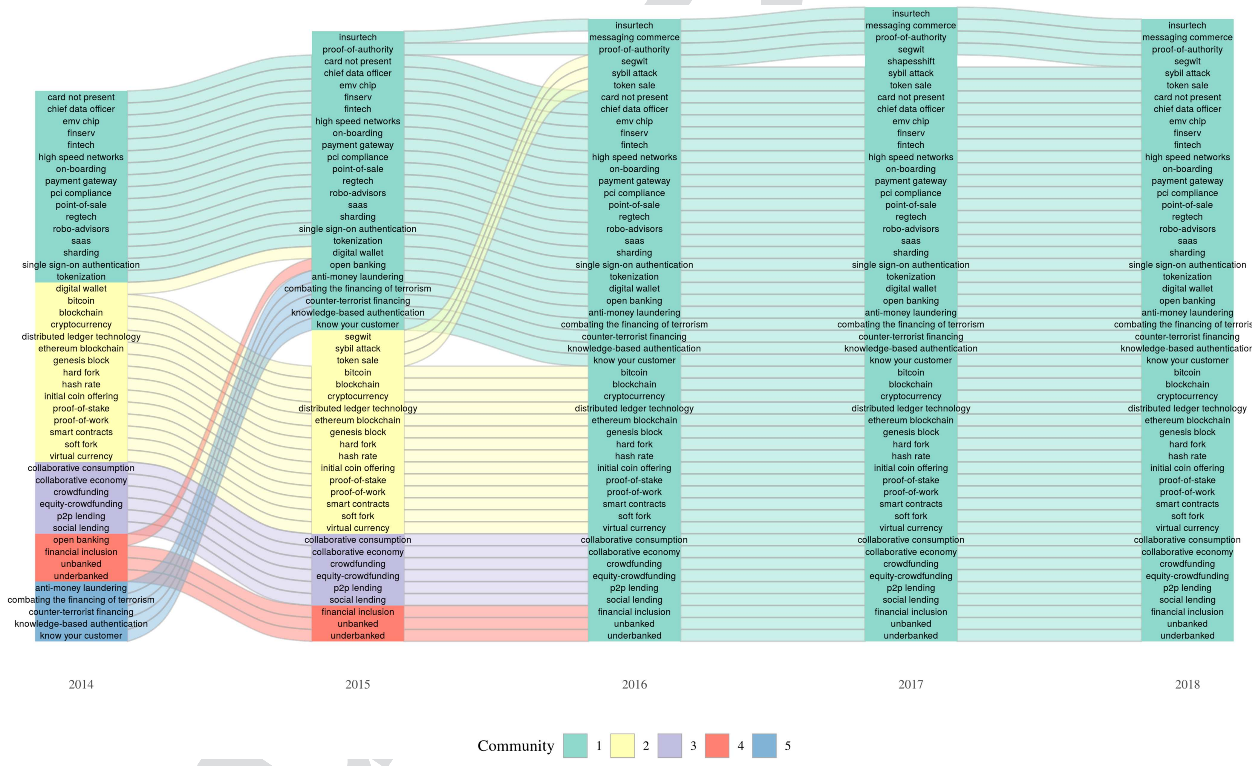

FIG. 2. Alluvial diagram of the communities found by the Infomap algorithm on the weighted co-occurrence networks of the years from 2014 (left) to 2018 (right). Each vertical set refers to a year. Colors are defining different communities. Fintech terms are shown in each box. We detect five communities in 2014, four communities in 2015, and one community from 2016 to 2018.

communities shown in Fig. 2. The algorithm detects five communities in 2014, four communities in 2015, and a single community starting from 2016. In summary, the weighted co-occurrence networks are becoming denser over time. We interpret the time evolution of the weighted co-occurrence network as the progressive setting of a coherent set of terms used in the business and technology area generically addressed as fintech. In Sec. IV, we will use this set of fintech terms to investigate the professional descriptions (written in the English language) of a large and heterogeneous set of companies operating in many countries.

IV. ANALYSIS OF PROFESSIONAL DESCRIPTIONS OF COMPANIES

In this section, we report on the analysis of fintech term occurrences detected in professional documents (i.e., documents written by economic analysts) describing the profile of companies operating in many countries. These are the descriptions of companies that are present in the Capital IQ database and in the Crunchbase Pro database. This set of professional texts is a relatively limited corpus comprising 2.2×10^6 documents.

We detect at least one term of the fintech list in 38 648 distinct descriptions of companies. We believe this number can be considered as a rough estimation of the number of companies currently focused on fintech. In fact, the number is about three times the estimate made by a McKinsey study in 2016.[13] In the last column of Table II, we report the occurrence of the 53 fintech terms in the documents of the dataset. Specifically, 50 out of 53 terms are detected in the documents describing the companies. The occurrence profile of the terms is pretty similar to the occurrence profile detected in the corpus of texts of news and blogs. In fact, the correlation between the occurrence of the 50 terms detected both in the texts of news and blogs and in the descriptions of companies is 0.824 (when measured as Pearson's correlation coefficient between term occurrence) or 0.891 (when measured as Spearman's correlation coefficient between term rank). This similarity of use of fintech terms in news and blogs and in professionally edited texts is another evidence supporting the assumption that the set of fintech terms defines a compact and coherent set of terms.

The databases have a number of attributes characterizing the companies. In the present study, we select country, municipality of the headquarter, and economic sector among them. A partial summary of these attributes is shown in Table IV. The table shows the 50 most common attributes of country (first and second column), economic sector (third and fourth column), and municipality (fifth and sixth column) with their occurrence. The table shows that the occurrence of all three attributes is heterogeneous. To provide a measure of the heterogeneity of occurrences we use the Herfindal index[14] that is a widespread simple measure of concentration of attributes of a set of elements. The Herfindhal index H of the reported attributes is $H = 0.223$ for countries, $H = 0.228$ for economic sectors, and $H = 0.0117$ for municipalities. High values of Herfindhal index indicate high concentration of the attribute in few elements, whereas low values indicate homogeneous distribution of the attribute to the different elements. The maximum value of the Herfindal index is one (complete concentration in one element). The minimum value of the Herfindhal index is equal to $H_{min} = 1/N_e$, where N_e is the number of elements. In the present case, the minimum value (perfect homogeneity) would be observed when $H_{min} = 0.006\,13$ for countries, $H_{min} = 0.0159$ for economic sectors, and $H_{min} = 0.000\,218$ for municipalities. The empirically observed values are all much above the values expected for homogeneous distributions of the attributes and indicate a high degree of heterogeneity. The heterogeneity of attributes reflects both the different diffusion of fintech interest and activities in different countries, municipalities and economic sectors and the heterogeneity of the databases discussed in Sec. II.

The bias of the databases and the heterogeneity of attributes make frequency analysis of the attributes not reliable. We, therefore, perform an over-expression analysis of the attributes observed in our datasets with a methodology used in network science. With this approach, we highlight over-expression of the presence of some fintech terms in the description of companies with different attributes of economic sector, country, and municipality of headquarters. This is achieved by selecting those pairwise relationships between an attribute of companies and fintech terms that cannot be explained by a null model of random connection that takes into account the heterogeneity of the attribute and of the fintech terms.

Let us comment in some detail the heterogeneity of the three investigated attributes. The country with the highest number of companies having fintech terms in their professional description is the United States. This is consistent both with the bias of databases (in the original set 61.3% of the companies are located in this country) and with the leading role that this country has in the fintech movement. However, in the set of companies having at least one fintech term in their description, the United States has 40.1% of the companies. This percent is still very high but less than the one observed in the original dataset. The United Kingdom has almost the same percent in the original (7.50%) and in the selected set (7.59%). A number of countries that we could label as innovative have higher percent in the selected set. For example, Canada has 1.51% in the original set and 4.78% in the selected set. Singapore has 0.378 % in the original set and 1.76% in the selected set. Israel has 0.244% in the original set and 1.17% in the selected set. Switzerland has 0.967% in the original set and 1.18% in the selected set. We interpret this change of the ranking as an indication that the databases are moderately less biased toward the United States and the United Kingdom when the coverage focuses on companies dealing with fintech topics, methods, or products. However, the bias is still quite strong and our analysis will explicitly take into account this limitation of the databases.

To characterize the economic sector, we use the industry classification of the Global Industry Classification Standard (GICS) developed jointly by Standard and Poor's and MSCI/Barra companies. GICS was developed in 1999 and it is periodically updated. The GICS structure today is organized in 11 sectors, 24 industry groups, 68 industries, and 158 sub-industries. In our analysis, we use the classification at the level of industries of July 2018. The 38 648 selected companies belong to 63 distinct GICS industries and the occurrence of the different industries is quite heterogeneous. In the third and fourth column of Table IV, we list the occurrence of the 50 most common industries. The heterogeneity of the industries is immediately evident. In fact, the most common industry *Internet Software and Services* is characterizing 13 891 companies, whereas

TABLE IV. Occurrence of the top 50 most common attributes of country (first and second column), economic sector (third and fourth column), and municipality (fifth and sixth column) of the companies presenting at least one fintech term in their company description. We also provide the total number of unknown for each type of attribute. Companies with at least one fintech term in their description belong to 163 countries, 63 industries, and 4474 municipalities.

Country	Occurrence	Industry	Occurrence	Municipality	Occurrence
United States	15 502	Internet Software and Services	13 891	London	1 720
United Kingdom	2 934	Software	8 582	New York	1 566
Canada	1 847	Capital Markets	3 729	San Francisco	1 216
China	1 317	IT Services	2 899	Singapore	669
India	1 237	Media	896	Paris	470
Germany	964	Professional Services	893	Toronto	457
France	907	Health Care Technology	646	Beijing	436
Australia	772	Electronic Equipment, Instruments and Components	559	Chicago	401
Singapore	680	Commercial Services and Supplies	502	Los Angeles	318
Switzerland	457	Diversified Financial Services	466	Boston	316
Israel	451	Banks	426	Berlin	307
Spain	434	Consumer Finance	364	Vancouver	291
Brazil	432	Technology Hardware, Storage and Peripherals	223	Austin	283
Netherlands	416	Insurance	149	Atlanta	279
Hong Kong	346	Real Estate Management and Development	126	Shanghai	279
Japan	304	Hotels, Restaurants and Leisure	124	Palo Alto	267
Ireland	291	Diversified Consumer Services	122	Mumbai	241
Italy	256	Diversified Telecommunication Services	106	Tokyo	231
Sweden	237	Internet and Direct Marketing Retail	95	Sydney	225
South Africa	229	Communications Equipment	91	Seattle	223
Russia	224	Containers and Packaging	81	San Diego	210
Finland	179	Healthcare Providers and Services	73	Dublin	205
Poland	175	Metals and Mining	68	Tel Aviv	181
South Korea	170	Distributors	47	Dallas	169
Denmark	159	Machinery	41	Amsterdam	168
Belgium	152	Trading Companies and Distributors	39	Denver	165
Mexico	145	Semiconductors and Semiconductor Equipment	34	Washington	159
New Zealand	144	Air Freight and Logistics	33	Melbourne	154
United Arab Emirates	127	Construction and Engineering	33	Miami	154
Austria	119	Wireless Telecommunication Services	33	Stockholm	153
Malaysia	118	Chemicals	31	San Jose	152
Estonia	117	Household Durables	31	Barcelona	151
Norway	106	Specialty Retail	29	Hong Kong	150
Indonesia	104	Thrifts and Mortgage Finance	27	Moscow	145
Argentina	101	Textiles, Apparel and Luxury Goods	26	Shenzhen	145
Nigeria	91	Electrical Equipment	25	Madrid	142
Turkey	87	Food Products	25	Mountain View	133
Philippines	84	Industrial Conglomerates	24	Menlo Park	132
Taiwan	82	Paper and Forest Products	22	Bangalore	130
Ukraine	79	Road and Rail	19	Seoul	128
Chile	73	Healthcare Equipment and Supplies	18	Munich	127
Portugal	70	Aerospace and Defense	16	Houston	122
Luxembourg	69	Biotechnology	14	San Mateo	121
Thailand	63	Beverages	12	Sao Paulo	119
Czech Republic	61	Food and Staples Retailing	12	Zug	116
Malta	56	Independent Power and Renewable Electricity Producers	10	Las Vegas	115
Lithuania	55	Life Sciences Tools and Services	10	Cambridge	113

TABLE IV. *(Continued.)*

Country	Occurrence	Industry	Occurrence	Municipality	Occurrence
Bulgaria	54	Oil, Gas and Consumable Fuels	10	Dubai	110
Cayman Islands	54	Airlines	6	Sunnyvale	110
Vietnam	50	Personal Products	6	Irvine	108
…	…	…	…	…	…
…	…	…	…	…	…
Unknown	4 479	Unknown	2 867	Unknown	5 280

the *personal Products* industry (50th in rank) is characterizing only six companies. The 18 industries with more than 100 occurrences belongs to 7 out of 11 sectors. Specifically, we have two *Industrials* (*Commercial Services and Supplies* and *Professional Services*), two *Consumer Discretionary* (*Hotels, Restaurants, and Leisure*, and *Diversified Consumer Services*), one *Health Care* (*Health Care Technology*), five *Financials* (*Capital Markets*, *Diversified Financial Services*, *Banks*, *Consumer Finance*, and *Insurance*), five *Information Technology* (*Internet Software and Services*, *Software*, *IT Services*, *Electronic Equipment, Instruments, and Components*, and *Technology Hardware, Storage, and Peripherals*), two *Communication Services* (*Media* and *Diversified Telecommunication Services*), and one *Real Estate* (*Real Estate Management and Development*). Even when we limit to sizable occurrences, the impact of the diffusion of fintech terms is on a broad number of economic sectors with a particular emphasis on Finance and Information technology. It is worth noting that the selected companies might be sometimes difficult to classify. In the above list of 18 top industries, three of them are classified by connoting them as "*Diversified.*" Moreover, the most frequent industry *Internet Software and Services* is described by analysts as "a relatively small industry primarily engaged in enabling and supporting commerce and other types of business transactions over the Internet. So, they offer cloud-based solutions and services that make customer interaction with businesses easier.[15] The definition of the industry within GICS was revised by Standard and Poor's and MSCI/Barra companies[16] at the end of 2018. Reclassification events are occurring in several areas and carry information about technological evolution.[17] Here, we interpret the reclassification event observed for the economic sector with the highest occurrence in the selected companies as an indication of the difficulty found by the analysts in defining nature and profile of the companies.

The third attribute we investigate is the municipality of the company location or headquarter. We have this information for 33 368 companies. They are located in 4474 distinct municipalities all over the world. The number of companies per municipality is again highly heterogeneous reflecting a Zipf like behavior.[18,19] In fact, when we regress the logarithm of the number of companies on the logarithm of the rank of the municipality, we obtain a power law exponent of -1.073 very close to the -1 value expected for a Zipf plot.

We observe a quite pronounced abundance of companies in some cities or metropolitan areas. The city with the largest number of companies is London UK. Other top cities are New York, San Francisco, and Singapore. In addition to San Francisco many other municipalities of the San Francisco Bay area are present in the top 50 municipalities (Palo Alto, San Jose, Mountain View, Menlo Park, San Mateo, Sunnyvale). By summing the number of companies operating in these municipalities of the San Francisco Bay area, one obtains 2131 companies, perhaps indicating the highest concentration of fintech companies in the world. Other metropolitan areas with a large number of companies are the great London area (1883 companies) and the New York City area (1738 companies). The list also contains small and medium size municipalities. One interesting example is the municipality of Zug in Switzerland having 116 companies (rank 46). The valley where this municipality of 120 000 inhabitants is located is called the "crypto valley" and has hosted *The Crypto Valley Blockchain Conference* in 2019. On the other hand, the over-expression of companies with headquarters in Zug might also be related to the fact that Zug is a tax heaven for companies and the detected over-expression might only manifest the tendency of some of the companies dealing with fintech terms to locate their headquarters in a municipality with fiscal advantage.

Heterogeneity, and most probably uneven coverage of companies across different countries, is, therefore, present for all three attributes. Our analysis will, therefore, use a methodology that is robust with respect to the presence of it. To properly deal with this heterogeneity, we analyze relationships between company attributes and fintech terms as bipartite networks and we then detect over-expressed relationships.

Specifically, we start our approach by constructing three bipartite networks. The first is a countries–fintech term network, where we aggregate all companies located in the same country; the second is an economic industries–fintech term network, where we aggregate all companies working in the same economic industry, and the third is a municipalities–fintech term network, where we aggregate all companies working in the same municipality. The first network is a bipartite network with 163 countries and 50 fintech terms. The number of links is 1651 and the link density is 0.203. The second network is a bipartite network with 64 industries and 50 fintech terms. It has 707 links and a link density equals to 0.221. The third network is a bipartite network with 4474 municipalities and 50 fintech terms. In the third network links are 10 893 and the link density is 0.048.

To highlight the over-expressed relationships between countries, industries, and municipalities with fintech terms, we detect over-expressed links on all three networks. This is done by using the methodology of statistically validated network.[9,10] The detection of a statistically validated network (SVN) works as follows. Let us consider an attribute a of companies, whose occurrence is N_a and a fintech term b whose occurrence is N_b. Let us define $N_{a,b}$ as the number of occurrences of fintech term b in documents of companies with attribute a and let us call the total number of documents N_t. With these definitions, the probability of observing X co-occurrences of the attribute a and fintech term b under a null hypothesis of random mixing is well approximated by the hypergeometric distribution[9]

$$H(X|N_t, N_a, N_b) = \frac{\binom{N_a}{X}\binom{N_t-N_a}{N_b-X}}{\binom{N_t}{N_b}}. \tag{2}$$

The probability of Eq. (2) allows to estimate a p-value $p(N_{a,b})$ associated with the empirical observation of $N_{a,b}$ co-occurrences or more of attribute a and fintech term b. In fact, the p-value is

$$p(N_{a,b}) = 1 - \sum_{X=0}^{N_{a,b}-1} H(X|N_t, N_a, N_b). \tag{3}$$

With this approach, one can associate a p-value to all links of the bipartite network linking nodes of attributes of set A and fintech terms of set B by performing a statistical test. It is worth noting that the test highlights the over-expressions with respect to a null hypothesis that takes into account the heterogeneity of the attributes. In other words, the relationships highlighted by the test are not necessarily the most frequent but rather the ones that violates the null hypothesis assuming random connections between heterogeneous attributes and fintech terms.

For each bipartite network, the number of statistical tests to perform is given by the number of links that are present in the bipartite network. This number is relatively high and for this reason a multiple hypothesis test correction is useful to avoid a large number of false positive. In the present investigation, we use the control of the false discovery rate (FDR) as multiple hypothesis test correction[20] and we set to 0.01 the value of the

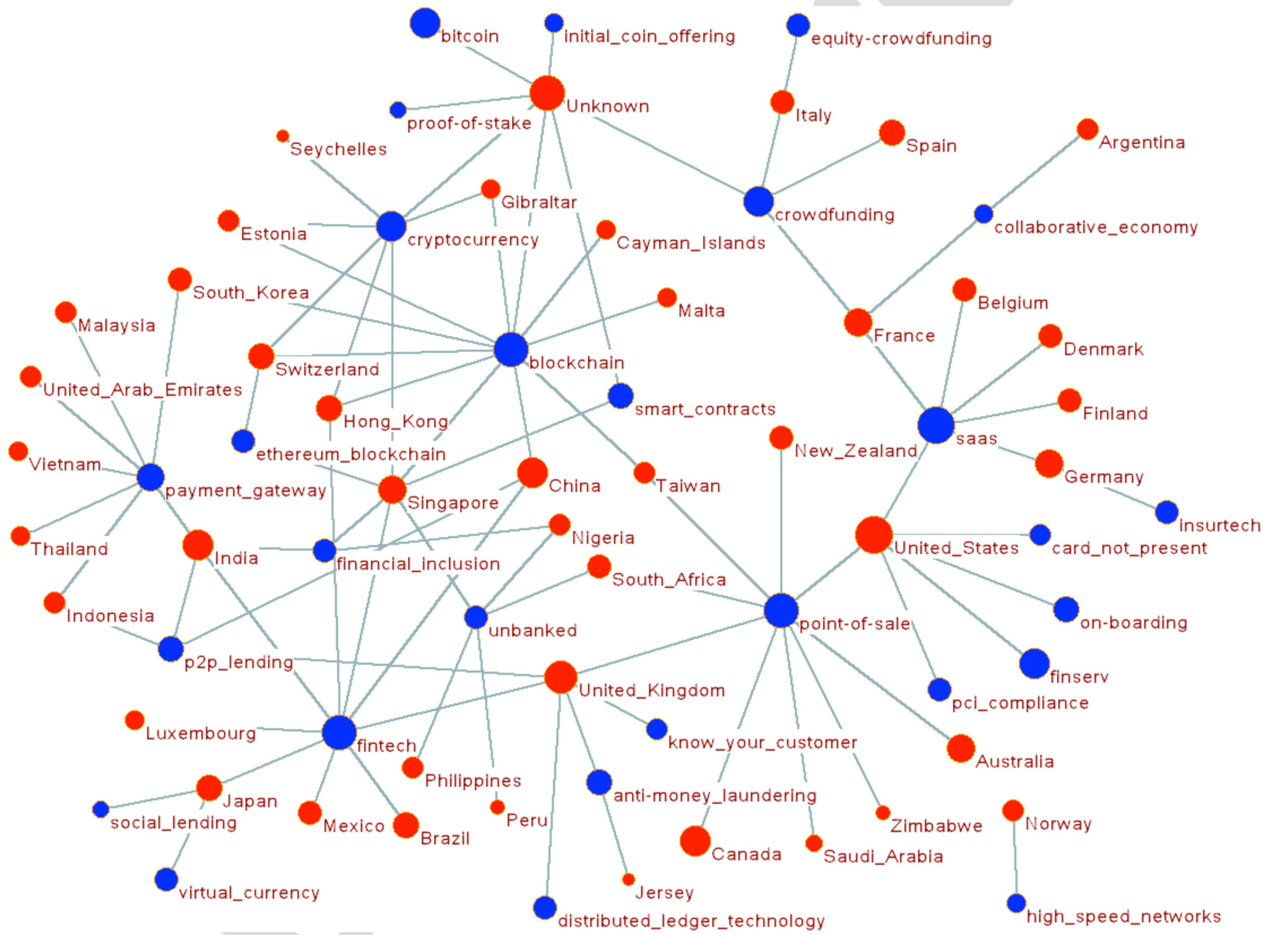

FIG. 3. Bipartite statistically validated network of countries–fintech terms. Blue nodes are fintech terms and red nodes are countries. For countries, the radius of each node is proportional to the logarithm of the number of companies of the country. For fintech terms, the radius of each node is proportional to the logarithm of the term occurrence.

false discovery rate, i.e., the expected maximal fraction of false positive.

We compute SVNs with a code written by us. However, programs computing SVNs from bipartite networks are available online.[21,22] Specifically, we have obtained SVNs of bipartite networks of (a) countries–fintech terms, (b) industries–fintech terms, and (c) municipalities–fintech terms.

The bipartite SVN of countries–fintech terms has 43 countries, 28 fintech terms, and 87 validated links. We are showing this network in Fig. 3. The blue nodes are fintech terms and the red nodes are countries. All the companies not reporting the information about the country in the databases are labeled by the term "Unknown." In the figure, the radius of each node describing a country (red nodes) is proportional to the logarithm of the number of companies of the country, whereas the radius of each node describing a fintech term (blue nodes) is proportional to the logarithm of the term occurrence.

By analyzing the figure, we note that countries where companies present an over-expression of the word *Blockchain* in their profiles are Gibraltar, Cayman Islands, Malta, Taiwan, China, Singapore, Hong Kong, Switzerland, South Korea, and Estonia. Mediterranean countries Italy, Spain, and France have companies over-expressed in *Crowdfunding* whereas north European countries Belgium, Denmark, Finland, and Germany present over-expression with *SAAS*. Germany has also an over-expressed link with *Insurtech*. Fintech terms *Unbanked* and *Financial inclusion* are over-expressed in companies of the following countries: India, Singapore, Nigeria, South Africa, Peru, and Philippines. All these countries except Singapore are developing countries with high potential of extension of financial inclusion.

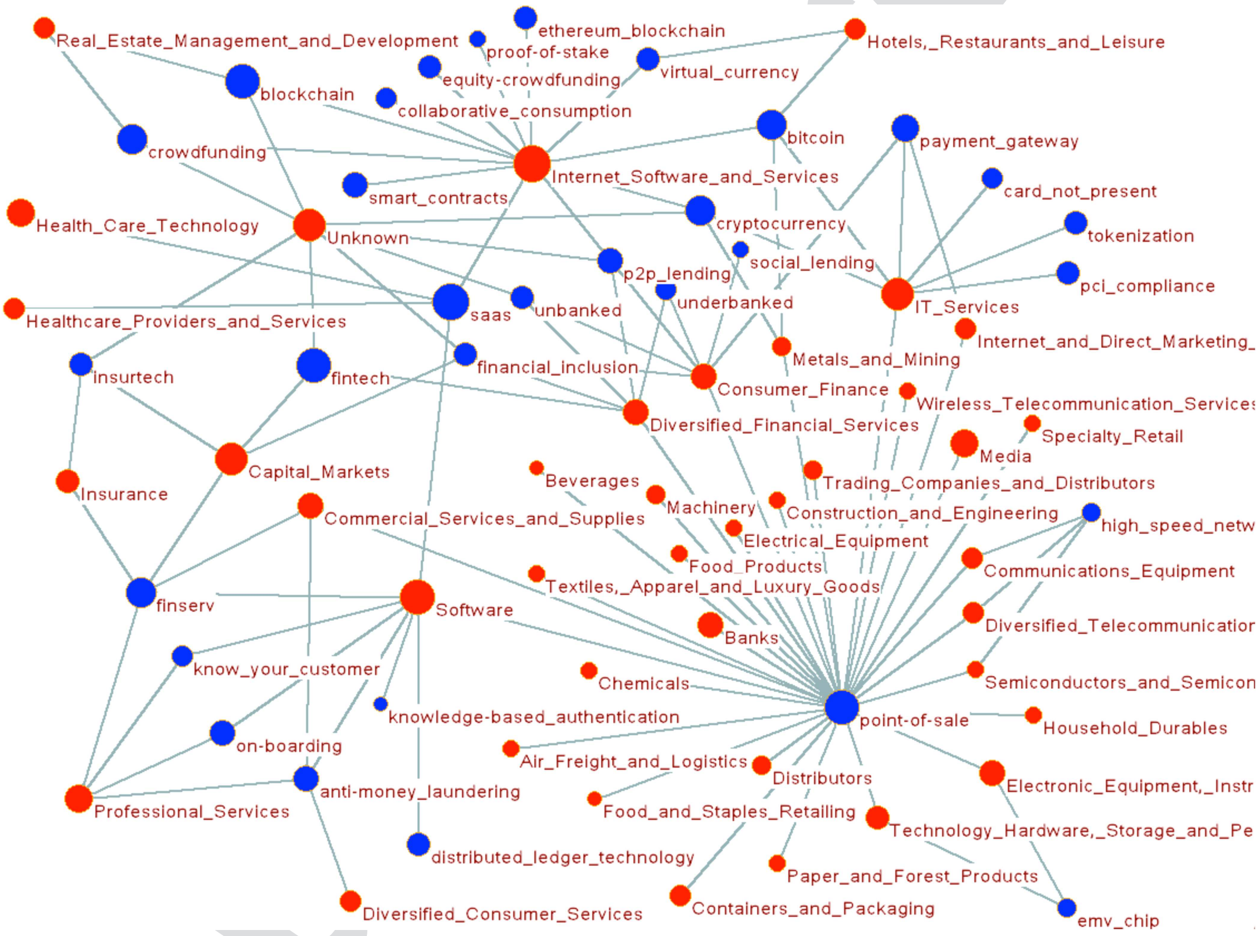

FIG. 4. Bipartite statistically validated network of industries–fintech terms. Blue nodes are fintech terms and red nodes are industries. For industries, the radius of each node is proportional to the logarithm of the number of companies of the industry. For fintech terms, the radius of each node is proportional to the logarithm of the term occurrence.

The bipartite SVN of industries–fintech terms has 40 industries, 31 fintech terms, and 101 validated links. The validated network is shown in Fig. 4. We note that the companies belonging to the *Internet Software and Services* present over-expression with some terms of the fintech list of terms. In fact, the companies of this industry are linked with *Blockchain*, *Collaborative consumption*, *Equity crowdfunding*, *Proof of stake*, *Ethereum blockchain*, *Virtual currency*, *Bitcoin*, *Cryptocurrency*, *P2P lending*, *SAAS*, *Smart contract*, and *Crowdfunding*. Companies of the industry of *IT services* present over-expressed links with the fintech terms of *Payment card industry (PCI) compliance*, *Tokenization*, *Card not present*, *Payment gateway*, *Bitcoin*, *Cryptocurrency*, and *Point of sale*. Companies belonging to the industry of *Software* or to the industry of *Professional services* present over-expressed links with *Finserv*, *Know your customer*, *On boarding*, and *Anti-money laundering*. Companies of the finance industries *Capital markets*, *Diversified financial services*, and *Consumer finance* are characterized by over-expression of the terms *Fintech*, *Finserv*, *Insurtech*, *Financial inclusion*, *Unbanked*, *Underbanked*, *P2P lending*, *Social lending*, *Payment gateway*, and *Point of sale*. It is also worth noting that several of the industries characterized by a limited number of companies (recognizable by nodes of small radius) are linked with *Point of sale*. Within fintech processes and services, this term is primarily used to address point of sale financing. Point of sales financing is the business practice allowing consumers to quickly finance large purchases with interest-free loans which are set up at the point of sale. Up until 2019, fintech firms have dominated this area.

The last bipartite SVN is the network of municipalities–fintech terms. The network detects 68 over-expressed links between 54 municipalities and 17 fintech terms. In Fig. 5, we show the network.

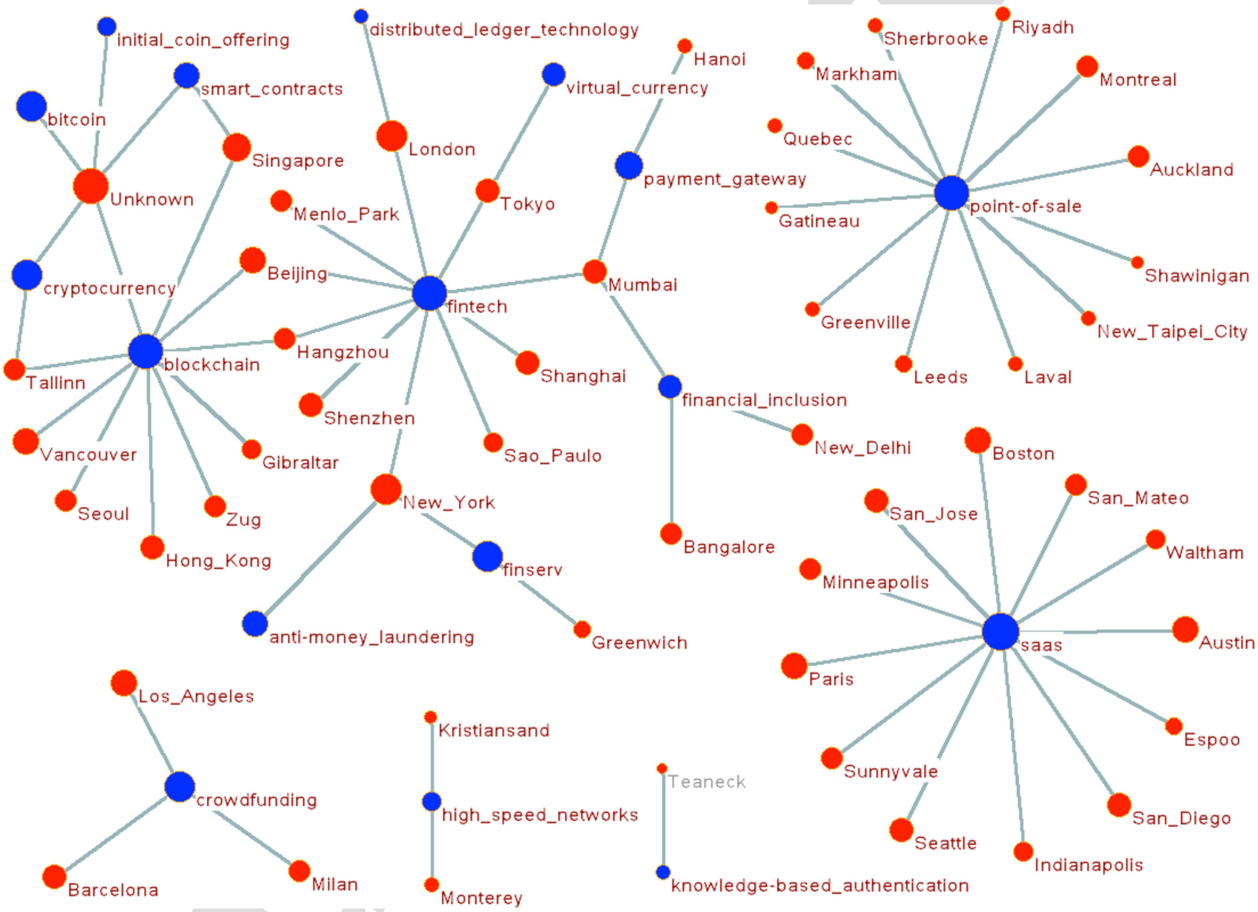

FIG. 5. Bipartite statistically validated network of municipalities–fintech terms. Blue nodes are fintech terms and red nodes are municipalities. For municipalities, the radius of each node is proportional to the logarithm of the number of companies of the municipality. For fintech terms, the radius of each node is proportional to the logarithm of the term occurrence.

In this case, the bipartite SVN shows several disjoint components. The largest component includes the fintech terms of *Cryptocurrency*, *Bitcoin*, *Initial coin offering*, *Smart contracts*, *Blockchain*, *Fintech*, *Distributed ledger technology*, *Virtual currency*, *Payment gateway*, *Financial inclusion*, *Finserv*, and *Anti-money laundering*. It involves cities that are hosting the biggest financial centers of the world such as *New York*, *Tokyo*, *Shanghai*, *Hong Kong*, *London*, *Shenzhen*, *Mumbai*, *Seoul*, and *Singapore*, and municipalities or cities with a strong tradition on digital innovation as *Menlo Park*, *Tallin*, and *Vancouver*. In the small municipality of Zug, companies present an over-expression of the term *Blockchain*, whereas the term *Financial inclusion* is over-expressed in companies located in *Mumbai*, *New Delhi*, and *Bangalore*. The other components of the network are characterized by a single fintech term. Specifically, these fintech terms are *Software as a service (SAAS)*, *Point of sale*, *Crowdfunding*, *High speed networks*, and *Knowledge-based authentication*.

V. DISCUSSION AND CONCLUSIONS

Our large scale textual analysis of news and blogs in the English language shows that a set of terms has developed and consolidated during the calendar years from 2014 to 2018 ending up in a compact and coherent set of terms used worldwide to describe fintech business activities. The search for this set of terms in the professional descriptions of a large dataset of companies located worldwide has faced the problem of the degree of coverage of databases in different countries. Databases are biased toward specific countries, and, therefore, a simple frequency analysis can be misleading. We, therefore, perform an analysis using a network science approach that is able to detect over-expression of a specific attribute with respect to a null hypothesis taking into account the heterogeneity of the investigated bipartite network.

With our approach, we obtain highlights about the over-expression of specific fintech terms in the description of a large number of companies of the fintech movement. Companies located both in developed and in developing economies present some degree of specialization (i.e., over-expression of occurrence of specific fintech terms in their professional description). Our analysis also shows that fintech topics, products, and services have the potential to impact a large number of industries. In fact, our analysis of the bipartite SVN economic sectors–fintech terms comprises 40 of the 63 economic sectors. One of the terms with several statistically validated links, *point of sale*, is also used outside the field of fintech. We have retained this term in our analysis because it plays an important role in the fintech business. In fact, point of sale financing is one of the main areas of development of fintech activities. By considering the use of the term *point of sale* outside fintech, we acknowledge that some of its links might not be uniquely related to point of sale financing. However, it is worth noting that the SVN approach is a pairwise approach and results obtained for a specific term do not affect results of other pairs. Therefore, in the unrealistic worst case that all links of *point of sale* term do not relate to point of sale financing, the remaining pairwise links between fintech terms and economic sectors would highlight over-expression of fintech terms in companies that are active in a minimum number of 22 distinct economic sectors.

We are also able to detect a geographical pattern of over-expression for companies dealing worldwide with fintech topics, services, and products. We characterize the geographical location down to the municipality of the headquarters of the companies. The over-expressions detected show that, in addition to the most important financial centers, a large number of companies are located in the San Francisco bay area and in a set of cities acting as innovation hubs of their countries. We are also able to highlight over-expression of small municipalities like Zug or Gibraltar that have clusters of companies with over-expression in the same area of the fintech business. Specifically, both municipalities have over-expression of blockchain in the descriptions of companies.

In summary, a methodology based on the analysis of bipartite networks constructed from biased or incomplete databases is able to highlight over-expressions of attributes of elements of the systems (in the present case companies). Our methodology is characterized by the control of false positives in the determination of statistically significant over-expressions. In other words, the over-expressions detected are all statistically significant at the chosen level of the control of false discovery rate ($\alpha = 0.01$). Unfortunately, a methodology simultaneously controlling the number of false positives and the number of false negatives is not yet available and, therefore, we cannot exclude a sizable number of false negatives.

In spite of this limitation, by relying on a full control of absence of false positives, our analysis unequivocally shows that fintech is a multi-industry, geographically distributed movement with a detectable level of geographical and economic sector specialization. This business movement is focusing on technical and methodological innovation of financial products, services, and activities. The innovations produced have the potential to deeply change the way mankind is dealing with finance in the coming years.

AUTHORS' CONTRIBUTIONS

F.C. and R.N.M. conceived the study. F.C. performed the text analysis of databases. F.C. and R.N.M. analyzed and interpreted the results and wrote the manuscript.

ACKNOWLEDGMENTS

We thank Luca Marotta for his help in preparing the alluvial diagram. R.N.M. acknowledges financial support of the project "Stochastic forecasting in complex systems" (Project No. 2017WZFTZP).

F.C. is employed by company Quid, San Francisco, CA, USA. R.N.M. declares no competing interests.

DATA AVAILABILITY

The data that support the findings of this study are available from Capital IQ, Crunchbase, and LexisNexis. Restrictions apply to the availability of these data, which were used under license for this study. Requests to access these datasets should be directed to Crunchbase https://about.crunchbase.com/products/crunchbase-pro/, LexisNexis https://www.lexisnexis.com/en-us/products/nexis/feature-get-the-story.page, and S&P Global (for Capital IQ) https://www.spglobal.com/marketintelligence/en/solutions/sp-capital-iq-platform.

REFERENCES

[1]Y. Elkana, *The Discovery of the Conservation of Energy* (Harvard University Press, 1974).

[2]P. Schueffel, "Taming the beast: A scientific definition of fintech," J. Innovation Manage. **4**, 32 (2016).

[3]D. W. Arner, J. Barberis, and R. P. Buckley, "The evolution of fintech: A new post-crisis paradigm," Environ. Innovation J. **47**, 1271 (2015). Q4

[4]A. Bettinger, "Fintech: A series of 40 time shared models used at manufacturers hanover trust company," Interfaces ■, 62–63 (1972).

Q5 [5]K. Börner, S. Sanyal, and A. Vespignani, "Network science," Annu. Rev. Inf. Sci. Technol. **41**, 537–607 (2007).

[6]M. Newman, *Networks: An Introduction* (Oxford University Press, 2010).

[7]A.-L. Barabási *et al.*, *Network Science* (Cambridge University Press, 2016).

[8]M. Á. Serrano, M. Boguná, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," Proc. Natl. Acad. Sci. U.S.A. **106**, 6483–6488 (2009).

[9]M. Tumminello, S. Micciche, F. Lillo, J. Piilo, and R. N. Mantegna, "Statistically validated networks in bipartite complex systems," PLoS One **6**, e17994 (2011).

[10]V. Hatzopoulos, G. Iori, R. N. Mantegna, S. Miccichè, and M. Tumminello, "Quantifying preferential trading in the e-mid interbank market," Quant. Finance **15**, 693–710 (2015).

Q6 [11]Online (last accessed April 28, 2019).

[12]M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," Proc. Natl. Acad. Sci. U.S.A. **105**, 1118–1123 (2008).

[13]D. Drummer, A. Jerenz, P. Siebelt, and M. Thaten, in *Fintech—Challenges and Opportunities: How Digitization is Transforming the Financial Sector* (McKinsey, Dusseldorf, 2016), pp. 1–7.

[14]S. A. Rhoades, "The Herfindahl–Hirschman index," Fed. Res. Bull. **79**, 188 (1993).

[15]S. Banerjea, "Internet-software and services outlook: Invest for the long haul," see https://www.zacks.com/commentary/180928/internet-software-services-out look-invest-for-the-long-haul (2018) (last accessed April 28, 2019).

[16]MSCI, "An announcement for the MSCI Global Standard Indices," see https://app2.msci.com/webapp/index_ann/DocGet?pub_key=AlQMAD0uvzk%253D%26lang=en%26format=html (2018) (last accessed April 28, 2019).

[17]F. Lafond and D. Kim, "Long-run dynamics of the US patent classification system," J. Evol. Econ. **29**, 631–664 (2019).

[18]G. K. Zipf, *The Psycho-Biology of Languages* (Houghton-Mifflin, Boston, MA, 1935).

[19]H. A. Simon, "Some further notes on a class of skew distribution functions," Inform. Control **3**, 80–88 (1960).

[20]Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," J. R. Stat. Soc.: Ser. B **57**, 289–300 (1995).

[21]C. Bongiorno, "Bipartite-tools," see https://github.com/cbongiorno/Bipartite-Tools (2017) (last accessed November 21, 2019).

[22]D. Challet, "Statistically validated networks," see https://cran.r-project.org/web/packages/SVN/index.html (2019) (last accessed November 21, 2019).