Understanding Ecosystem Data

Rahul C. Basole, PhD
Accenture AI
rahul.basole@accenture.com

Abstract

There is a growing body of empirical studies on business ecosystems. Driven by different questions these studies typically employ a wide variety of data sources – ranging from open to proprietary, structured to unstructured – that contain a broad range of entities, relationships, activities, and issues of interest. Individually, these data sources offer the ability to investigate very targeted business ecosystem questions. However, when linked and combined these data sources can potentially open up many new lines of inquiry. The purpose of this study is to provide an overview of the scope and complexity of the business ecosystem data landscape, discuss what type(s) of information is captured in them, identify how data sources overlap and differ, discuss strengths and weaknesses, and suggest new types of analyses that could be generated when combined. In doing so this study aims to help researchers and practitioners with the data identification and selection process and stimulate novel data-driven ecosystem intelligence. The study concludes with theoretical and managerial implications.

1. Introduction

In a highly complex, dynamic, and global business environment, it is increasingly important to have a better understanding of the structure and dynamics of the various ecosystems a firm is embedded in [1, 2, 3]. Robust and timely ecosystem intelligence can lead to earlier anticipation of risks and opportunities [4], improved competitive benchmarking [5, 6], and better strategic decision-making [7, 8, 9, 10, 11, 12, 13, 14, 15].

One of the key elements for achieving effective business ecosystem intelligence is access to relevant, high-quality, and comprehensive business ecosystem data [3, 16]. Firms frequently leverage diverse internal information systems to make sense of their broader enterprise ecosystem. However, these sources often lead to myopic insights, as many important external signals about partners, competitors, markets, and other contextual matters can be missed [17].

To overcome this issue, firms often turn to external data to augment their existing ecosystem intelligence sources. External data sources can provide a richer, expanded, and complementary view of competitive activities, developer interactions, product offerings, and supply chain relationships [18, 13, 4, 19]. In fact, as even external data becomes more common, firms are looking for alternative data to achieve further competitive intelligence differentiation.

Integratively, these demands, along with increased digitization and access advancements, continue to fuel the enormous growth of the data-as-a-service space. New players focusing on specific data segments are continuously emerging while existing data providers are seeking innovative ways to differentiate their offerings through new partnerships or business models. Yet, it is exactly this scale, scope, and speed of business data sources that can be challenging to make sense of. What relevant ecosystem data sources exist? How do they differ? What types of questions can be answered with these data sources? And what is the quality and cost of these data sources?

A comprehensive analysis of the data landscape for ecosystem intelligence does not exist to the best of my knowledge. Ecosystem scholars use a variety of methods to curate their data for their analysis. When contexts are relatively new or unknown, researchers frequently use primary data collection methods, such as case studies or interviews [9]. While proven to be valuable for theoretical and conceptual development, however, these methods do not scale well when trying to understand large complex ecosystems. A growing number of scholars have started using secondary data sources to advance our understanding of ecosystems. Sources generally included established data providers such as Bloomberg or Thomson Reuters. Typically these data sources are only available at significant license fees and potentially only affordable for larger institutions.

URI: https://hdl.handle.net/10125/64444
978-0-9981331-3-3

With the advent of the digital economy, however, we have seen a significant increase in socially-curated, open datasets. Indeed, a lot of ecosystem intelligence can now be derived from these publicly available data sources. When appropriately mined and curated and combined, important insights can be gained.

The objective of this study is multifold. First, I want to identify as comprehensively as possible the data sources currently available for ecosystem analysis. Some of these data sources are well known and used extensively in scholarly studies. Others are relatively new and likely unknown to many researchers and practitioners. By exposing them, I hope to draw attention to alternate sources that could be used. To guide my thinking, I develop a conceptual ecosystem data framework. Second, I want to describe and analyze the commonality and differences between these data sources. I pursue this objective by using a data-driven "competitor" approach. The resulting map will uncover the complexity of the ecosystem data landscape and guide researchers in selecting potentially unique and complementary data sources. Third, I will discuss issues associated with ecosystem data curation and use that researchers should be aware of. Lastly, I propose potential research directions and questions that could be pursued when leveraging the panoply of data sources. Cumulatively, I hope that this study will stimulate future data driven ecosystem research beyond singular and commonly used data sources to studies that fuse and combine data sets. I conclude the study with implications and extensions.

2. Framing the Ecosystem Data Landscape

Figure 1 provides a conceptual summary of the different considerations I made in framing the ecosystem data landscape. In the sections that follow, I will discuss each of them in detail.

2.1. Ecosystem Primitives

Before I begin identifying potential data sources for ecosystem analysis, it is pertinent to understand what types of ecosystem primitives (or building blocks) researchers are typically interested in. At a high-level, ecosystem research is interested in the configuration of different types of entities (i.e., "nodes") and their connections (i.e., "links").

At the node level, the most common examples include companies, people, products and services, patents, and technologies, which are often further differentiated by types and categories, such as incumbents companies and startup ventures, suppliers and customers, executives and board members, patent applications and granted patents, and so on.

At the link level, researchers have examined a wide range of direct connections between entities, including interfirm (e.g., R&D, supply chain, marketing, licensing, technology transfer, joint ventures, alliances), investments, acquisitions, or mergers, or represent intra-firm hierarchies, like parents organizations, business units, or subsidiaries. Links could also include inter-personal relationships, patents citations, or activities (like launches, litigations, etc.). Moreover, connections can be derived or computed from content or association, such as the co-occurrence on boards or similarity between products. Generally, relationships are either undirected, directed, or bidirectional.

2.2. Data Source Types

Broadly considered, there are two types of data: *unstructured* and *structured* data.[1] The term unstructured generally refers to the fact that no identifiable structure within this data is available. Others describe unstructured data as data that cannot be stored in rows and columns in a relational database. Typical examples of unstructured data thus includes textual documents and fields, audio, images, and videos. While no classification is needed for unstructured data, controlled navigation is also not possible. Unstructured data is generally very rich in content/information, but also can have a low signal to noise ratio without appropriate techniques. Common techniques to analyze unstructured text documents include full-text search, text mining, or natural language processing. Video and image based files can be analyzed using computer vision approaches. On the other hand, structured data follows a predefined schema (i.e., conforms to some specification). A typical example of structured data is a relational database, in which content is created, assigned, and populated into a schema, which defines the type and structure of data as well as its relations. Most commonly used ecosystem data sources are structured (for ease of querying, identification, and presentation).

2.3. Data Source Categories

With an understanding of typical ecosystem primitives and types of data, we can now delineate broad data source categories in which the relevant data may exist. Existing ecosystem research has leveraged

[1] It can be argued that there are really three types of data sources, unstructured, fully structured, and semi-structured. In this study, however, we do not make the differentiation between structured data sources.

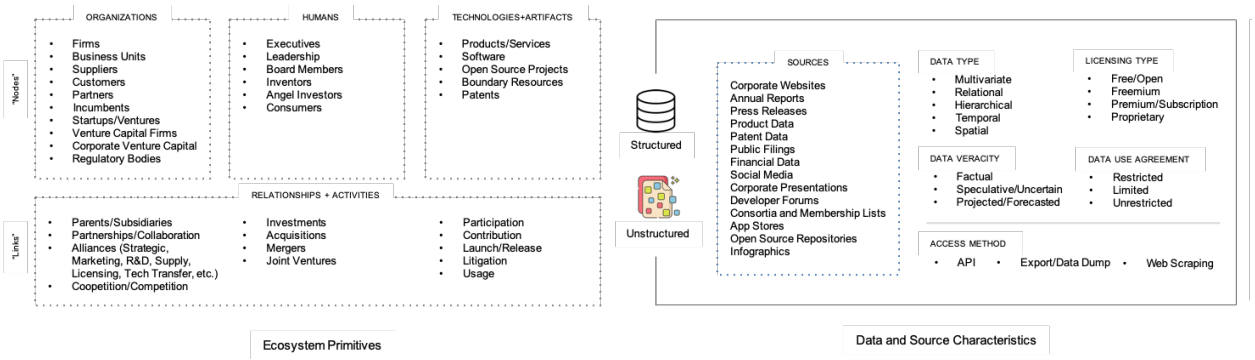

Figure 1. A Conceptual Ecosystem Data Framework.

a wide range of data source categories for each of the ecosystem primitives. For "organization" related data, for instance, studies have leveraged corporate websites to identify lists of partners, suppliers, and customers [2]. Organizational data is generally present in all data sources, including a wide range of unstructured data, such as press release, annual reports, public filings, and social media. "Human" related information, such as executives, board members, or investors, for instance is also frequently available on corporate website or professional networking sites. Developer activities can be extracted from developer forums and open source repositories as well as knowledge exchanges. "Technology and Artifact" related data can be found in patent data, app stores, and open source repositories, just to name a few. Some sources focus on specific types of artifacts, like APIs, software stacks, or new applications. Similarly, artifact and technology information can be found in press releases, blogs, and industry reports. "Relationship and activity" related data is quite pervasive in each of these data sources as well but generally is the most challenging to curate and organize. There are some data sources that identify specific relationships types curated from public filings (e.g., alliances). Most activity and relationship information, however, is either embedded in textual content and needs to be derived or inferred.

All of the different ecosystem primitives are likely embedded in audio, video, and image based data as well. For instance, corporate presentations and infographics often contain visual information about competitors and customers as well as artifacts (e.g., products and services). Talks can contain narrated information about ecosystem entities as well. Customer sentiment, public narratives about individuals, companies, and artifacts are often found in social media.

2.4. Data Types

Understanding data types is critical for determining the types of ecosystem analyses the data can afford as well as what ecosystem modeling approaches should be used. Moreover, understanding the underlying data type can help in selecting appropriate visual encodings when ecosystems are graphically represented.

Broadly considered, ecosystem data sources contain a wealth of multivariate, relational, hierarchical, temporal, or spatial data for different ecosystem primitives. At the organizational level, it could include the type of organization, the date it was founded, acquired, or dissolved, where it is located, who their customers and suppliers are, or what business units or parent organizations it contains. Similar data could be found for human and technology/artifact ecosystem primitives. All of these data types are also contained in unstructured data sources, such as textual documents, but need to be identified and extracted. For instance, annual reports could include forward and backward looking statements (temporal information) and mentions of investors and competitors (relational).

2.5. Data Veracity

At this point it is important to note that the veracity of data can vary significantly across data source types and categories. The veracity can range from factual (in most structured data sources) to speculative, projected, forecasted, or uncertain (in unstructured data sources). Of course, even data contained in structured data sources could have some veracity issues and needs be validated and verified. As researchers pursue ecosystem research, an understanding of the level of data veracity can help with the analysis and interpretation of the results. For instance, if data has low veracity, conclusions drawn from the study must be taken with a grain of salt. Later I will discuss some of the implications of low data

veracity and what ecosystem researchers could do to overcome this.

2.6. Licensing Types

Data providers utilize a wide range of licensing models to provide end users access to their data assets. These range from proprietary, subscription, and paid to freemium, free, and open. Each of these licensing options has implications for the researcher, in particular in terms of the resources needed to obtain them.

2.7. Data Use Agreement

Closely linked to licensing types is the concept of data use agreements. While publicly available data sources are of tremendous value, how the data can be used will be delineated by these agreements. In general, providers encourage scholarly research based on their data. If used for commercial purposes, proper legal vehicles need to be put in place. Often times when the data source is proprietary, paid, or premium, the data provider allows the end user to use it only for specific purposes. Ecosystem researchers must be aware of the fineprint of these usage agreements (contained in the FAQ or legal sections) and email the data provider for any clarification and permission. This will protect the researcher from any subsequent legal consequences.

2.8. Access Methods

In order to utilize the data sources, the data needs to be made available in a useful format. Broadly considered there are three types of data access methods, including APIs, export/downloads/data dumps, and web scrape. Each of these methods is generally linked to the data licensing models as well as data use agreements. Many freemium and proprietary data providers make it easy for researchers to obtain data in a useful format, either through APIs, export/download capabilities, and in rare occasions full data dumps. The quality and extent of the APIs can vary greatly, with some having very rich data extraction capabilities while others are more limited. Often times, the number of calls that can be made is linked to the subscription/licensing price. Ecosystem researchers need tom keep that in mind when deciding on a source as well as on a data extraction/curation method. When neither of the vendor-provided access methods are available, researchers can utilize web scraping methods. Of course this needs to be done with the data use agreements in mind. According to recent legal rulings, however, publicly facing information can be scraped and utilized. There are many ways to conduct web scraping, with some off-the-shelf tools available.

3. Methodology

Considering our conceptual ecosystem data framework, there are unquestionably a multitude of data sources that could be used for ecosystem analysis tasks. Moreover, it is clear that no single data source could likely be used to answer all questions. The scale and diversity of data sources is also reflected by the diverse empirical contexts examined in the rapidly growing business ecosystem literature. Past work has argued that the choice of data is generally guided by the underlying question(s) or task(s) the investigator has (e.g., [1]). It is probably also reasonable to assume that new data sources will continue to emerge on a regular basis. As such, I do not suggest that the list identified in this study will be complete. However, to generate the most comprehensive, relevant list of data sources, I used a multiple perspective identification approach, focused on the broad ecosystem data characteristics I delineated in the data landscape framework.

First, I generated an extensive list of data sources and their providers I used or was broadly familiar with.[2] Next, I identified all prior work on ecosystems and reviewed each study for the type of data used by the authors. To do so, I leveraged existing literature reviews and meta analysis (e.g., [9, 20, 21, 22]) and used a forward/backward snowballing method to identify relevant studies. After eliminating non-empirical papers and those that did not use a secondary data source, I recorded the corresponding data source and provider. Third, I used a web-based search engine to identify additional relevant ecosystem data sources and providers, using a variety of keyword searches (e.g., “data sources”, “competitive intelligence data”, “company data”, “global market and economic data”). Lastly, I used the list of sources identified in Step 1-3 to find related/similar offerings through Google’s “People Also Searched For” capability. I developed custom webscrape scripts to iteratively identify related companies/offerings.

These data curation steps led to the identification of 107 core data source providers (excluding the generic categories of corporate website, venture capital websites, and consortia/membership websites). For each of these data sources, I explored their website to understand the value and content of the data source and identify the ecosystem characteristics identified earlier.

Since our goal is to understand the structure of the

[2]This is an important distinction as some data sources are synonymous to the data provider (e.g., Crunchbase). In other instances, a single company may offer multiple data assets, such as Thomson Reuters (e.g., SDC Platinum, VentureXpert, etc.).

ecosystem data landscape, it was pertinent to determine an approach to identify similar/dissimilar data sources. There are several different paths to achieve this. In this study I chose the competitor criteria approach. In other words, if a data source is listed as a competitor to another data source, I deemed it to contain similar information. In the spirit of ecosystem research, I utilized two publicly available data sources (Owler[3] and AlternativeTo[4]) to identify and cross-validate the list of competitors for each of our 107 core firms. This led to identification of 495 additional organizations, for a total of 602 data providers. I converted this competitor data into a data ecosystem graph, in which nodes represent data providers and edges represent whether two data vendors are deemed competitors. While the information can be directional (data provider A can be a competitor of data provider B but not necessarily vice versa), I treated this as an undirected graph, scaling the edges proportionally by the number of edges (*max*=2).

Finally, I used Gephi 0.92 to visualize the corresponding graph [23]. I used the force-directed Yifan Hu multi-level layout algorithm with default parameters [24] to position nodes. A force-based layout is based on the idea that network entities are shaped by mechanical laws, assigning repulsive forces between nodes and attraction forces between endpoints of edges. The use of a force-based layout is particularly appealing when the motivating issue is to identify central or prominent nodes, peripheral actors, or clusters in an ecosystem. To ensure readability and aesthetics, I followed several visual design principles, including no node overlap and edge crossing minimization. The node size is proportional to a data vendors's importance in the ecosystem as measured by its Betweenness centrality. To gain insight into the presence of subcommunities, I color encoded with the corresponding modularity class. I used a NoOverlap algorithm to space out nodes and address potential visual occlusion issues.

4. Results

Table 1 provides a summary of the core 107 data sources including an availability of ecosystem primitives (Organizational (Org); Human (Hum); Technology and Artifacts (T+A); Relationship and Activities (Rel+Act).

Figure 2 shows the full ecosystem data landscape map. Overall, the figure shows a moderately interconnected network with a clear core and several peripheral clusters. Clusters that are disconnected from the main component represent sources that may be "distant" from commonly utilized sources. The graph contains 602 nodes and 1002 edges (main component contains 563 nodes (92.6%) and 952 edges (95.01%). The visualization reveals that incumbent data providers are located in the core, such as S&P Global Market Intelligence, Thomson Reuters, Bloomberg, and Dow Jones (provider of Factica), focused on organizations. Closely positioned to this cluster are data sources associated with startup and ventures such as Pitchbook, Crunchbase, PrivCo, CB Insights, DataFox, and Owler. Near the core are also a range of "human" data providers, such as Glassdoor and Indeed, who are connected to the core by LinkedIn. Located more at the periphery are technology and artifact data providers, such as BuiltWith, AppAnnie, Apptopia, SimilarWeb, and AlternativeTo. Interestingly, patent data is almost fully disconnected from all other data sources.

5. Discussion

5.1. Data Identification

Our analysis has hopefully revealed the enormous breadth and depth of available data sources. However, simply having access to data does not equate to ecosystem intelligence. Effective data-driven ecosystem decisions requires identification of relevant data. But what does "relevance" really mean in a business ecosystem intelligence context? This is a rather difficult question to answer without an understanding of the underlying decision making context. At one extreme, decision makers ideally would want all possible data necessary (i.e., the proverbial "kitchen sink") to answer a particular question or task. However, this can be a daunting and resource-intensive endeavor given the plethora of data sources. To avoid unnecessary data curation processes, drawing appropriate system boundaries corresponding to the ecosystem analysis tasks is thus critical. The choice of boundaries is driven by the nature and intent of the problem, the questions being asked, and the costs involved [1]. In many instances, boundaries can be determined a priori. Examples include specific market segments, entity characteristics, geographies, or date ranges. However, often times the boundary is fluid and emerges over time, particularly when the intelligence analysis is exploratory in nature. Effective business ecosystem intelligence tools must thus provide the flexibility to acquire and integrate additional and/or new data when needed.

5.2. Data Linking

The central tenet of our study is that ecosystem intelligence tasks likely require information that is found across multiple data sources. In order to

[3] https://www.owler.com/
[4] https://alternativeto.net/

Table 1. Core Ecosystem Data Sources/Providers

#	Source	Org	Hum	T+A	Rel+Act	#	Source	Org	Hum	T+A	Rel+Act	#	Source	Org	Hum	T+A	Rel+Act
1	Aberdeen	X	X		X	37	FullContact	X	X		X	73	Product Hunt	X		X	X
2	AlternativeTo	X		X	X	38	G2 Crowd	X	X		X	74	ProgrammableWeb	X		X	X
3	AngelList	X	X			39	GitHub	X	X	X	X	75	Quandl	X			X
4	App Annie	X		X	X	40	GitLab		X	X	X	76	Radius	X			X
5	Apptopia	X		X	X	41	Glassdoor	X	X		X	77	RainKing	X	X		X
6	Barchart			X	X	42	Google	X	X	X	X	78	RelPro	X	X		X
7	Bazaarvoice		X		X	43	GrowthIntel	X	X		X	79	RelSci	X	X		X
8	Bitbucket		X	X	X	44	Hoovers	X	X		X	80	Reuters	X	X	X	X
9	Bloomberg	X	X	X	X	45	IBISworld	X	X		X	81	S&P Global Market Intelligence	X	X	X	X
10	Bloomberg Tradebook				X	46	Indeed	X	X		X	82	Salesloft	X	X		X
11	BoardEx	X	X		X	47	Infogroup	X	X		X	83	SeedInvest	X			X
12	Bombora	X	X		X	48	Innography	X	X	X	X	84	SelectHub	X		X	X
13	Builtwith			X	X	49	InsideView	X	X		X	85	SEMrush			X	X
14	Bureau van Dijk	X	X		X	50	IPlytics	X		X	X	86	Siftery	X		X	X
15	CareerBuilder	X	X		X	51	Knoema	X	X		X	87	SimilarWeb	X		X	X
16	CB Insights	X	X	X	X	52	LeadIQ	X	X		X	88	Slideshare	X	X	X	X
17	Cigital			X	X	53	LeadsAI	X	X		X	89	Spiderbook	X	X	X	X
18	Comscore	X	X		X	54	Leadspace	X	X		X	90	StackOverflow		X	X	X
19	CoRepo	X	X		X	55	LexisNexis	X	X		X	91	Startup Ranking	X			X
20	Corporate360	X	X		X	56	LinkedIn	X	X		X	92	Techcrunch	X	X	X	X
21	Crunchbase	X	X	X	X	57	Macrobond	X	X		X	93	Thomson Reuters	X	X	X	X
22	Data.com	X	X	X	X	58	Mattermark	X			X	94	Tracxn	X	X		X
23	Datafox	X	X	X	X	59	Mergermarket	X	X		X	95	Trustpilot	X			X
24	Dealflow	X	X		X	60	Mintel	X	X		X	96	TrustRadius				X
25	Demandbase	X	X		X	61	Morningstar	X	X		X	97	Twitter	X	X	X	X
26	Digimind	X		X	X	62	MSCI	X			X	98	USPTO			X	X
27	DiscoverOrg	X			X	63	Nasdaq	X			X	99	VC Experts	X	X		X
28	Dow Jones	X	X	X	X	64	Northern Light	X	X	X	X	100	Vertical Knowledge	X			X
29	Duedil	X	X	X	X	65	NPD	X		X	X	101	Wefunder		X		X
30	Dun & Bradstreet	X	X		X	66	Owler	X	X		X	102	Wikipedia	X	X	X	X
31	Dun & Bradstreet Credibility Corp	X			X	67	PatSnap	X			X	103	Xignite	X	X		X
32	Enigma	X	X		X	68	Payscale	X	X		X	104	Xing	X	X		X
33	Euromonitor International			X	X	69	PitchBook	X	X		X	105	ZipRecruiter	X	X		X
34	F6S	X		X	X	70	Plainsight Intel	X			X	106	Zirra	X	X	X	X
35	Facebook	X	X		X	71	Preqin	X			X	107	ZoomInfo	X	X	X	X
36	FactSet	X	X		X	72	PrivCo	X	X		X						

make these data sources actionable, data linking is needed [25, 26]. Data linking (also referred to as duplicate identification, record linkage, data matching, entity resolution or reference reconciliation) is arguably one of the most critical steps in business ecosystem intelligence. Data matching involves the identification and matching of relevant ecosystem entities (e.g. organizations, people, products, technologies, etc.) within and across data sources. In a perfect world, data sources would have clear mechanisms to be integrated and combined. Despite advances in computational data resolution techniques, ranging from deterministic to probabilistic, human intervention is often necessary [27]. Without proper matching capacities, significant data quality problems arise. Consider the matching of company names. While many company identifiers exist, they are not uniformly applied across data sources. Companies use different business type abbreviations, change names, or get acquired. Moreover, many companies have subsidiaries and operate under different names nationally. Variations can thus range from subtle to substantial. If this seems complicated, data matching is even more complicated when dealing with people, products, and technologies. By combining multiple different external data platforms, organizations can generate significant novel and triangulated ecosystem insights.

5.3. Data Quality and Governance

"Garbage in is garbage out." The well-known adage certainly applies to ecosystem intelligence as well. In order to achieve effective ecosystem intelligence it is pertinent to ensure that the quality and integrity of the data is high. Most structured data sources often have very rigorous data curation processes in place to ensure that the underlying quality of the data is right. Often times, data is inspected, validated, and verified by a team of experts and non-experts. Socially-curated data sources leverage the wisdom of the crowd to identify, edit, and manage the quality of data. Others use algorithmic techniques to compare their data with other datasets to identify and mitigate gaps.

An important aspect for ecosystem researchers using secondary data sources is to understand how their data sources of choice are curated and updated. Even if the data is provided by an established provider it is important to question and probe the quality of the data by comparing or triangulating the information with other data sources. As we identified above, almost all data sources have an alternative. Ideally, we would use them integratively to compare content to establish data veracity. As data is updated and evolves it is important to understand the underlying data lineage, what transformations were made.

The quality of unstructured data requires similar approaches, albeit there are potentially more difficult to pursue. As I mentioned above, data veracity is an

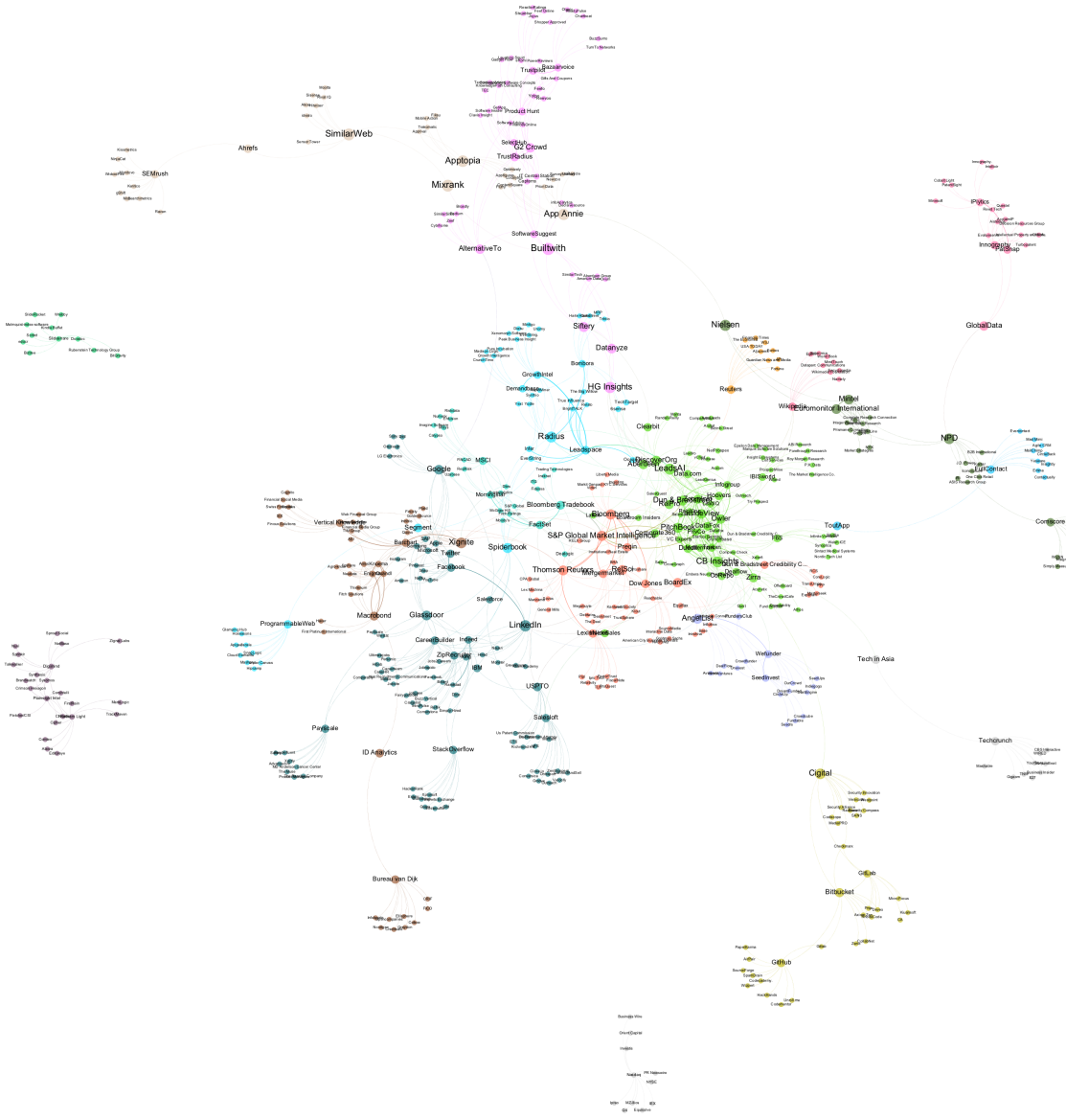

Figure 2. Ecosystem Data Landscape.

issue and can have an impact on data quality. Structured data often capture factual and realized information. Investments that have been made, patents that have been filed, leadership changes that have taken place, products that have been launched. All of these things are somewhat retrospective data points and can be verified and validated. Unstructured data, however, can contain forward looking information that have a component of uncertainty and ambiguity associated with. For instance, annual statements, analyst reports, and blogs could speculate about the nature of relationships or entities. When applying text analytics to such data, it is pertinent to understand the nuances of such statements. This is not an issue of data quality but more an advice to ensure that adding a qualification of the nature of the data (factual, uncertainty, ambiguous) may be rather prudent.

5.4. Data Halflife

It can be argued that ecosystem intelligence data has a halflife (i.e., the average timespan it is valuable to a decision maker). The value of data diminishes based on the cadence of decisions. Ecosystem decision tempos range from tactical (real-time) to operational (hours-weeks) to strategic (quarterly-years). If relevant data is not used in the appropriate timeframe, its value and usefulness diminishes quickly. Consequently, data, like radioactive material, has a halflife. While many business ecosystem intelligence decisions will have long timeframes, most organizations need to utilize a blend of data tempos for different tasks. Acquisition or investment decisions, for instance, are more long term and may require an evaluation of historic data. Dynamically adjusting supply chains to changing conditions (e.g. weather) clearly demand real-time data. Organizations that understand the halflife of their data, match it with their ecosystem intelligence tasks, and adjust their decision making tempo accordingly are most likely going to create competitive advantages.

5.5. Data Derivatives and Mashups

The data landscape I have discussed so far has largely considered each data source as a self-contained, independent entity with some data alternatives. However, these are not the only sources that can exist. In fact, there are some sources that have filtered down datasets (e.g., Startup Data Trends) and elements for the purpose of a particular analysis context (e.g., IEN Data [28]). Even established data providers are recognizing the need to integrate with other data sources. Crunchbase, for instance, has created a data marketplace integrating their core startup data with data from other vendors. We refer to these new data sources as data derivatives and mashups.

6. Ecosystem Research Opportunities

The identification of potentially new data sources and a mapping of their similarity provides the basis for exploring new ecosystem research opportunities. In particular, I argue that four pertinent opportunities can and should be pursued: triangulated insights, multi-level analyses, what, why, and context questions, and supply and demand analysis.

6.1. Triangulated Insights

Arguably one of the most beneficial aspects of using multiple linked data is triangulated insights and perspectives. Data sources that provide similar information can be used to check, verify, and validate data against each other. In doing so, data quality and validity can be improved. At the same time, multiple data sources can also help provide alternate perspectives into a phenomenon. For instance, a company's activities in a particular technology space (e.g., AI) could be examined through the lens of a firm's investment activities, its patenting behavior, and its collaboration with ecosystem players. Together, these datasets could provide temporal insights and also explain strategic behaviors and activities.

6.2. Multi-Level Analysis

Most ecosystem research has focused on a single level/unit of analysis, such as at the firm-, people-, or technology level. However, with linked data at different levels of an ecosystem interesting multi-level research can be pursued. Consider the following three scenarios and the new insights one may be able to derive (see Figure 3). Combining startup (organization), API (artifact), and investor (organization/people) data, we can explore the boundary resource footprint of emerging ventures and identify venture capitalists who are focused on investigating in the foundation of the digital economy. Linking developers, code repositories, and organizations we can examine the degree of company participation in open source projects and the scale and extent of the developer base. Connecting patents to technologies and organizations we could identify innovators and their footprint in emerging technologies, but also identify dark nets like dependencies and litigations. There is no question that I use an oversimplification to present these cases but I believe they provide the glimpse into the possibilities. The value of simply connecting two data sources is potentially greater than the use of a single source. What

needs to be cautioned here, however, is that if more ecosystem primitives are included in the analysis the interpretability of the results could become challenging. New models and methods needs to be developed to analyze, visualize, and make sense such research contexts.

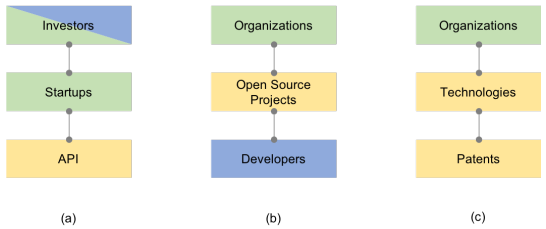

Figure 3. Examples

6.3. What, Why, and Context Questions

One of the common critiques of empirical research is that we make simplifying assumptions and often limit the variables that are included. By analyzing unstructured data, researchers have the ability to provide a contextual background to what and why something happened. Data-driven evidence can provide the needed contingent explanations to distill the behavior and outcomes we observe. The data can be used quantitatively (if coded appropriately) or qualitatively to substantiate findings.

6.4. Supply and Demand Side Analysis

Another exciting opportunity for using multiple data sources is to provide both supply and demand side insights into an ecosystem. In part this is closely related to the triangulated insights benefit mentioned above. For instance, researchers could use job postings to determine the skills and expertise companies are seeking in a fast-pacing technological environment and connect this to developer activities in open source code repositories and knowledge exchanges. Similarly, we could examine the evolution of software stacks of companies and the investments technology vendors are making in different solutions.

7. Concluding Remarks

The primary purpose of this study was to provide an overview of the scope and complexity of the business ecosystem data landscape, discuss what type(s) of information is captured in them, identify how data sources overlap and differ, discuss strengths and weaknesses, and suggest new types of analyses that could be generated when combined. In doing so, I hope to have highlighted the issues and stimulated researchers to think about ecosystem research and possible opportunities.

As I stated at the onset of this study, it is highly likely that I have missed or not included a dataset/data source. I have tried to the best of my knowledge to identify all data sources that have have been used in prior empirical ecosystem research, determine promising alternatives to these data sources, and propose entirely new data sources that have recently emerged. I view the list I have generated as a living document, which I hope ecosystem researchers will contribute to and enhance over time. I fully expect that new data sources will be added over time.

An obvious limitation is that my analysis of data sources focused on general ecosystem data sources and not any domain specific datasets. Researchers interested in particular domains are likely to find very targeted websites and data sources for them. For instance, to understand the video game ecosystem, researchers could leverage sites like Gamespot or IGN, which provide information on products, developers, console makers, and customers. Similar sites exist in other domains like energy, finance, or retail. We also did not consider the wealth of blogs or industry analyst reports provided by market research firms.

My hope is that this paper will trigger a rich dialog and perhaps lead to a community of interest in regards to ecosystem data. The links to all data sources as well as their characteristics will be made available as a living document. I am considering two avenues to do this. One opportunity is to create a Wikipedia entry that the community contributes to and maintains. Another would be to create a dedicated website that could act as a social-curated ecosystem research resource.

In the future, I would like to provide additional information for each data source, including the licensing models and access methods.

Our visualization of related data sources utilizes a search co-occurrence approach. While this is a valuable proxy in determining the similarity between data sources, more sophisticated approaches can and should be used in the future. One approach could involve the development of a feature vector of all salient characteristics and the computation of the distances between them. The results could then be plotted using a MDS approach. Another could simply count the number of overlapping characteristics. Each of these approaches could provide differentiated insights.

References

[1] R. C. Basole, M. G. Russell, J. Huhtamäki, N. Rubens, K. Still, and H. Park, “Understanding business ecosystem

dynamics: A data-driven approach," *ACM Transactions on Management Information Systems (TMIS)*, vol. 6, no. 2, p. 6, 2015.

[2] B. R. Iyer and R. C. Basole, "Visualization to understand ecosystems," *Communications of the ACM*, vol. 59, no. 11, pp. 27–30, 2016.

[3] R. C. Basole, J. Huhtamäki, K. Still, and M. G. Russell, "Visual decision support for business ecosystem analysis," *Expert Systems with Applications*, vol. 65, pp. 271–282, 2016.

[4] R. C. Basole, "Visual business ecosystem intelligence: Lessons from the field," *IEEE computer graphics and applications*, vol. 34, no. 5, pp. 26–34, 2014.

[5] R. C. Basole, "Visualization of interfirm relations in a converging mobile ecosystem," *Journal of information Technology*, vol. 24, no. 2, pp. 144–159, 2009.

[6] P. C. Evans and R. C. Basole, "Revealing the api ecosystem and enterprise strategy via visual analytics," *Communications of the ACM*, vol. 59, no. 2, pp. 26–28, 2016.

[7] R. Adner, "Ecosystem as structure: an actionable construct for strategy," *Journal of Management*, vol. 43, no. 1, pp. 39–58, 2017.

[8] M. G. Jacobides, C. Cennamo, and A. Gawer, "Towards a theory of ecosystems," *Strategic Management Journal*, vol. 39, no. 8, pp. 2255–2276, 2018.

[9] K. Järvi, S. Kortelainen, *et al.*, "Taking stock of empirical research on business ecosystems: a literature review," *International Journal of Business and Systems Research*, vol. 11, no. 3, pp. 215–228, 2017.

[10] M. Lyman, R. Ref, and O. Wright, "Cornerstone of future growth: Ecosystems," 2018.

[11] D. J. Teece, "Business ecosystem," *The Palgrave Encyclopedia of Strategic Management*, pp. 1–4, 2016.

[12] P. J. Williamson and A. De Meyer, "Ecosystem advantage: How to successfully harness the power of partners," *California management review*, vol. 55, no. 1, pp. 24–46, 2012.

[13] R. C. Basole, M. A. Bellamy, and H. Park, "Visualization of innovation in global supply chain networks," *Decision Sciences*, vol. 48, no. 2, pp. 288–306, 2017.

[14] M. Iansiti and R. Levien, "Strategy as ecology," *Harvard business review*, vol. 82, no. 3, pp. 68–81, 2004.

[15] R. Kapoor, "Ecosystems: broadening the locus of value creation," *Journal of Organization Design*, vol. 7, no. 1, p. 12, 2018.

[16] J. Huhtamäki, M. G. Russell, and K. Sill, "Processing data for visual network analysis," *Visual Analytics for Management: Translational Science and Applications in Practice*, p. 56, 2016.

[17] R. C. Basole, A. Srinivasan, H. Park, and S. Patel, "ecoxight: Discovery, exploration, and analysis of business ecosystems using interactive visualization," *ACM Transactions on Management Information Systems (TMIS)*, vol. 9, no. 2, p. 6, 2018.

[18] R. C. Basole, T. Clear, M. Hu, H. Mehrotra, and J. Stasko, "Understanding interfirm relationships in business ecosystems with interactive visualization," *IEEE Transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 2526–2535, 2013.

[19] M. A. Russell, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. " O'Reilly Media, Inc.", 2013.

[20] L. Thomas and E. Autio, "Modeling the ecosystem: a meta-synthesis of ecosystem and related literatures," in *DRUID 2012 Conference, Copenhagen (Denmark)*, 2012.

[21] K. Manikas and K. M. Hansen, "Software ecosystems–a systematic literature review," *Journal of Systems and Software*, vol. 86, no. 5, pp. 1294–1306, 2013.

[22] P. K. Senyo, K. Liu, and J. Effah, "Digital business ecosystem: Literature review and a framework for future research," *International Journal of Information Management*, vol. 47, pp. 52–64, 2019.

[23] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in *Third International AAAI Conference on Weblogs and Social Media*, 2009.

[24] Y. Hu, "Efficient, high-quality force-directed graph drawing," *Mathematica Journal*, vol. 10, no. 1, pp. 37–71, 2005.

[25] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data: The story so far," in *Semantic services, interoperability and web applications: emerging concepts*, pp. 205–227, IGI Global, 2011.

[26] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," *Synthesis lectures on the semantic web: theory and technology*, vol. 1, no. 1, pp. 1–136, 2011.

[27] P. Christen, *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[28] N. Rubens, K. Still, J. Huhtamäki, and M. G. Russell, "Leveraging social media for analysis of innovation players and their moves," *Innovation Ecosystems Network, Media X, at Stanford University*, 2010.