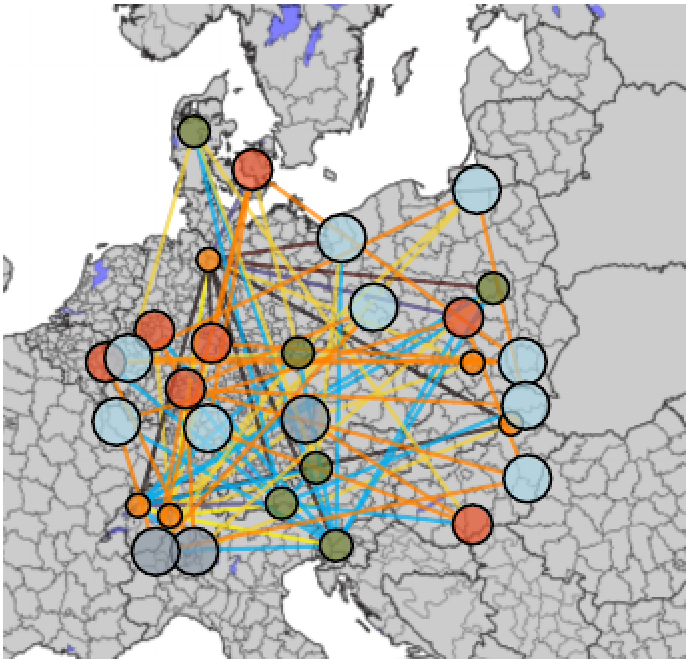
Open Risk White Paper

Connecting the Dots: Concentration, diversity, inequality and sparsity in economic networks

Authors: Philippos Papadopoulos

June 30, 2021



www.openriskmanagement.com

The open future of risk management

Contents

1 SUMMARY

In this second Open Risk White Paper on *Connecting the Dots* we examine measures of concentration, diversity, inequality and sparsity in the context of economic systems represented as network (graph) structures. We adopt a stylized description of economies as property graphs and illustrate how relevant concepts can represented in this language. We explore in some detail data types representing economic network data and their statistical nature which is critical in their use in concentration analysis. We proceed to recast various known indexes drawn from distinct disciplines in a unified computational context.

1.1 Structure of the white paper

- The introduction places the white paper in the context of contemporary academic research and information management technologies and motivates the need for an integrated approach that enables taping into these distinct knowledge domains.
- The section on **network data** summarizes the technical machinery used in representing economic networks as graphs. The core concept of a property graph and the data structures it supports are described in the detail required for discussing concentrations and constructing the associated measures.
- The section on **index construction mechanics** discusses the main choices and calculation steps required to transform property graph data sets into concentration measurements.
- Finally, the **index catalog** section enumerates a large number of the most commonly used indexes, categorized according to their broad nature.

1.2 Further Resources

- The concentrationMetrics is an open source framework written in Python that allows the easy calculation of many of the measures discussed in this paper.
- The Open Risk Manual is an open online repository of information for risk management developed and maintained by Open Risk. Various concentration risk concepts mentioned in this White Paper are further documentd and explained using dedicatedOpen Risk Manual entries.
- The Open Risk Academy offers a range of online courses around risk and portfolio management, which utilize the latest in interactive eLearning tools, including a dedicated category on Credit Concentration Courses and Concentration Measurement Using Python
- More content: Open Risk White Papers and Open Risk Blog

1.3 About Open Risk

Open Risk is an independent provider of training and risk analysis tools to the broader financial services community. Our mission is captured by the motto: The open future of risk management. Learn more about our mission and projects at: `www.openriskmanagement.com`

2 Introduction

2.1 Context

Economic networks of interconnected agents engaging in exchange of goods, services and contracts of various forms are the defining attribute of human economies. While this is intuitively obvious, the quantification and analysis of economic networks has not played an important role in the historical development of economic and financial theory and practice. Yet in recent times, supported by developments in information technology both academics and practitioners connect the dots by analyzing economic phenomena making use of *graph theory* and *network analysis*. Such developments help create more detailed understanding of the structures and interactions between economic agents, ultimately guiding towards better policies and risk management.

Examples of economic networks already studied include for example the system of interconnected financial institutions. Potentially adverse linkages were an issue that rose to prominence after the financial crisis of 2008 (e.g., [1],[2] and a recent review in [3]).

In broader economic context, networks show up e.g. in core/periphery networks ([4]), the structure of multinational affiliate networks ([5]) and shadow banking ([6]). The ambition to understand the economy as a living network provides important motivation for collecting relevant statistics ([7]) and underpins modern flow-of-funds analysis ([8]). Empirical analysis is crucial, for example, to understand how interlinked households and non-financial firms affect borrower delinquencies and defaults which ultimately affects the health of the entire system. ([9], [10]). The explicit modeling of credit contracts as networks, in particular involving the banking system is now a fruitful line of inquiry ([11], [12], [13]).

In many disciplines outside finance and economics there are also growing bodies of work that make use of generalized graph structures as representations of complex networks. These are going under a variety of names: *multilayer networks*, *multiplex networks* or *multidimensional graphs*. Good overviews (and links to open source software that enables working with such structures) are given in ([14],[15]).

Turning to the focus of this paper, *concentration measurement* is one of the most important and frequent risk management analyses and is deeply linked to measurement concepts such as quantifying inequality, sparsity or diversity. Thinking and developing suitable such measures has a long historical tradition:

> *”It is generally agreed that, other things being equal, a considerable reduction in the inequality of incomes found in most modern communities would be desirable. But it is not generally agreed how this inequality should be measured”.*

A hundred years after the above quote, which comes from the seminal Dalton paper [16] on measuring inequality it is not clear if we are closer to a clear answer. There is an enormous literature on constructing and selecting appropriate methods for measuring inequality through indexes. In parallel, there is large and ever growing set of proposals and approaches covering concentration, diversity, sparsity or clustering as those phenomena are studied in other disciplines. Many methods have been introduced and used extensively within their respective domains (See Box 1 for a summary of the main disciplines).

In a previous paper [17] we reviewed the definitions of widely used concentration metrics such as the concentration ratio, the HHI index and the Gini and clarified their meaning and relationships in a probabilistic context. This analytic framework helped clarify the apparent arbitrariness of simple concentration indexes and brings to the fore underlying unifying concepts behind these metrics, thereby enabling their more informed use in portfolio and risk management applications. Expanding the scope of concentration

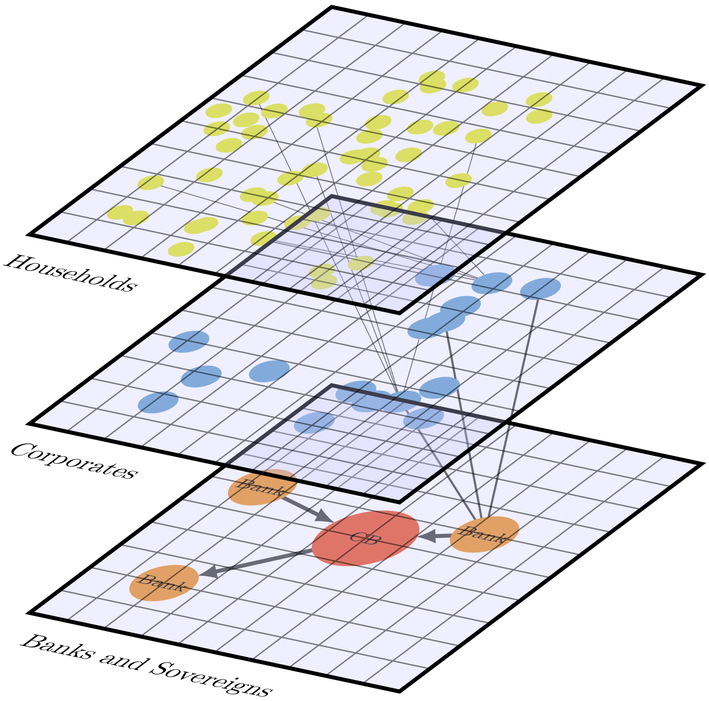


Figure 1: A multi-layer network representation of a stylized economy. Nodes are legal entities (real persons, firms, banks etc). Connections are indicating economic transactions or contractual relations.

risk analysis we ask the question: *How do the objectives and tools of concentration risk analysis translate in the context of economic networks?*

This question frames the main discussion topic of this white paper. To explore the question we continue developing the quantitative framework first introduced in [18]). The approach stylizes the description of economic networks as *contractual relationships* between agents described as *property graph* (Fig. 2.1). The term "property graph" is used primarily in the context of modern database systems ([19], [20]) and places an emphasis on the *storage* of information (versus mathematical graph properties and algorithms).

In the next section we will summarize the general concepts and notation that are part of the *property graph representation* of economic networks[1] but for brevity only do so as required for discussing concentration metrics. A more detailed discussion that focuses on detailed descriptions of contracts and balance sheets of economic agents is given in [18]).

2.2 Defining concentration measurement

Simply put, what underlies concentration or diversity or sparsity or inequality measurement is the assessment of the degree to which a particular property or properties are distributed across an extended system. For the concept of *distribution* to be important the system we study must contain multiple instances of similar components (for example a large number of economic actors belonging to the same category). It is implied that (at least in principle) a diverse set of possible distribution profiles are possible and the system studied is but one configuration out of many.

[1] We will use the term graph and network as synonymous

A general quantification concept that is a common building block is *statistical dispersion.* Indeed statistical moments such as variance or kurtosis are frequently used in concentration analysis. From physics and computer science one can also import concepts such as entropy (measuring the degree of system order or disorder).

The use of the word *measurement* suggests that the objective of concentration measurement is to focus preferentially on quantitative information and tools. **Qualitative considerations** creating concentration might be very important, especially for systems that are not adequately studied and quantified. We assume here that the concentration issues drawing our attention have had at least a preliminary expression in quantitative form.

Concentration is most simply expressed as a single number (outcome), with a defined range and (ideally) a set of indicative threshold values the define concentration levels. Such a number is usually called an *index, measure or metric.* Quite frequently indexes bear the name of the first authors proposing them in academic literature.

Concentration indexes are, quite generally, *mathematical maps* (functions) from some multivariate statistical (probability) distribution of observed values into a single real number (the index). They are thus essentially *summarizing* the content of a very complex object (that is a potentially higher dimensional) to a manageable (communicable) scalar value.

As mentioned already, concentration indexes can be considered as a type of *summary statistic or descriptive statistic* and indeed several approaches are based on dispersion measures (second or higher moments) of the distributions. In more formal treatments the different families of indexes might be segmented according to which *axioms* they are required to satisfy. There are many dozens of indexes proposed over the years and there are many more minor variations.

Despite commonalities in the four domains mentioned (concentration, diversity, inequality, sparsity) and associated measurement tasks, there are fairly significant differences in motivation, intended use and typical datasets for each domain. There is also great diversity in terminology and conventions which may obscure otherwise similar concepts. For simplicity we will use the term ”concentration index” as an overall category.

There are in particular possible axioms (requirements) of how a concentration measure should behave (e.g. when combining two systems) that are relevant in some domain but not useful in others. Thus the requirements and constraints for concentration measurement of economic network data depend on the context and need not be universal.

Box 1. The many guises of concentration measurement

The concept of concentration measurement is important both in academic research and in practical applications in a number of diverse disciplines. The specific nature of each field leads to important *field-specific adaptations*:

- *Inequality Measures*: Inequality indexes are constructed and studied in sociology, economics and policy work. This is maybe the most developed domain from a theory perspective as it incorporates utility preferences and selects suitable functional forms on the basis of rigorous axiomatic frameworks. On the other hand the focus is on important but rather specific numerical variables such as income and wealth distributions.
- *Concentration Risk Measures*: Concentration enters in various areas of finance and economics when assessing industial, market share concentration, market risk or credit portfolio risk concentration. A wide range of approaches is employed in practice: from simple numerical measures to sophisticated simulation based risk measures that are only computable through Monte Carlo simulations. This integration of concentration metrics with modeled *risk measures* is rather unique in this domain.
- *Sparsity Measures*: In signal processing sparsity means that a small number of (spectral) coefficients contains a large proportion of the energy. The concept of sparsity is very flexible and is also applicable in machine learning. This is possibly the most context-agnostic application domain, focusing on an information theoretic assessment without additional constraints or insights.
- *Diversity Measures*: These are common tools in ecology when assessing biodiversity. Diversity indexes focus on species abundances (thus mainly categorical data) instead of numerical data that are more common in inequality and concentration analysis. A unique aspect of biodiversity measures are *multi-scale* considerations.
- *Spatial Concentration*: In various domains utilizing geospatial data there is a need for indexes that express spatial concentration. Here the concentration measurement is intrinsically *multi-dimensional*. This introduces an expanded toolkit which aims among others to identify *spatially close* entities.
- *Clustering and Centrality Measures*: In network theory a distinct category of metrics aims to characterise the clustering of network connections in a graph. This domain too, requires highly specialized tools to extract usable information from graph structures.

Practically all of the above can be repurposed in the context of economic networks, providing a rich tapestry form characterising the distribution of properties of such networks.

3 Economic Network Data

In this section we will summarize a technical machinery (data structures) that can be used for representing economic networks. The core concept we will use is that of a *property graph.* Before we discuss it let let us briefly revisit the simple *mathematical graph.*

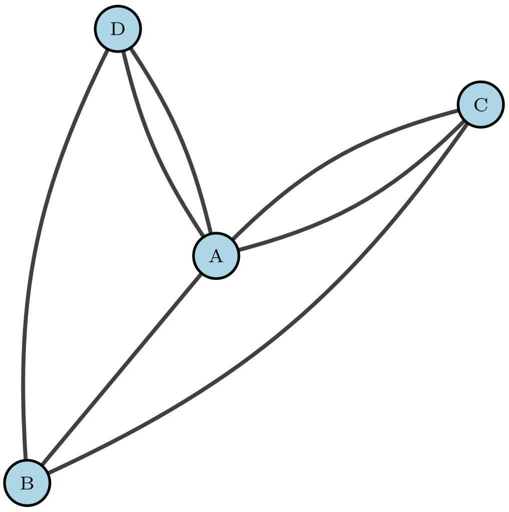
The classic (simple, undirected) graph is an ordered pair $G = (V, E)$ comprising of:

- V, a set of vertexes (also called *nodes*)
- $E \subseteq \{(x, y)|(x, y) \in V^2 \wedge x \neq y\}$, a set of edges (also called *links*), which are unordered pairs of vertexes (i.e., an edge is associated with two distinct vertexes).

In our context, nodes will typically be the economic entities (for example persons or legal entities) that we want to model (represent). Edges will be a natural tool to express *economic relationships* between those entities. Such relationships can be any concrete economic fact: transactions, contracts, shareholding interests etc.

As an example, in the abstract graph below four nodes are depicted (A, B, C, D). There are also some links between them (some nodes are linked in multiple ways).

Example of a simple graph



The classic graph as introduced above is an expressive and flexible mathematical tool. For example one can prove powerful theorems about graph properties that are generally valid across any network. But in terms of data modeling tools it has obvious limitations that must be addressed for many practical applications. E.g., what is the meaning of nodes and links? It could be anything, e.g. who-knows-whom, who transacts with whom etc. but in most practical situations additional information is required.

Historically the next level of *fidelity* in graph theory has been the introduction of *valued graphs.* A valued graph is a graph where a real number (value) is assigned to each edge. This generalization opens up additional representation possibilities. The edge value might be e.g., the credit exposure of one entity to an other, yet this is also far from sufficient. In addition in many concrete analyses one needs additional amounts of data to describe nodes (e.g. size, type etc.)

3.1 Property Graphs

In order to capture a reasonably realistic amount of economic network information we need to expand the mathematical structure of the valued graph in various non-trivial ways:

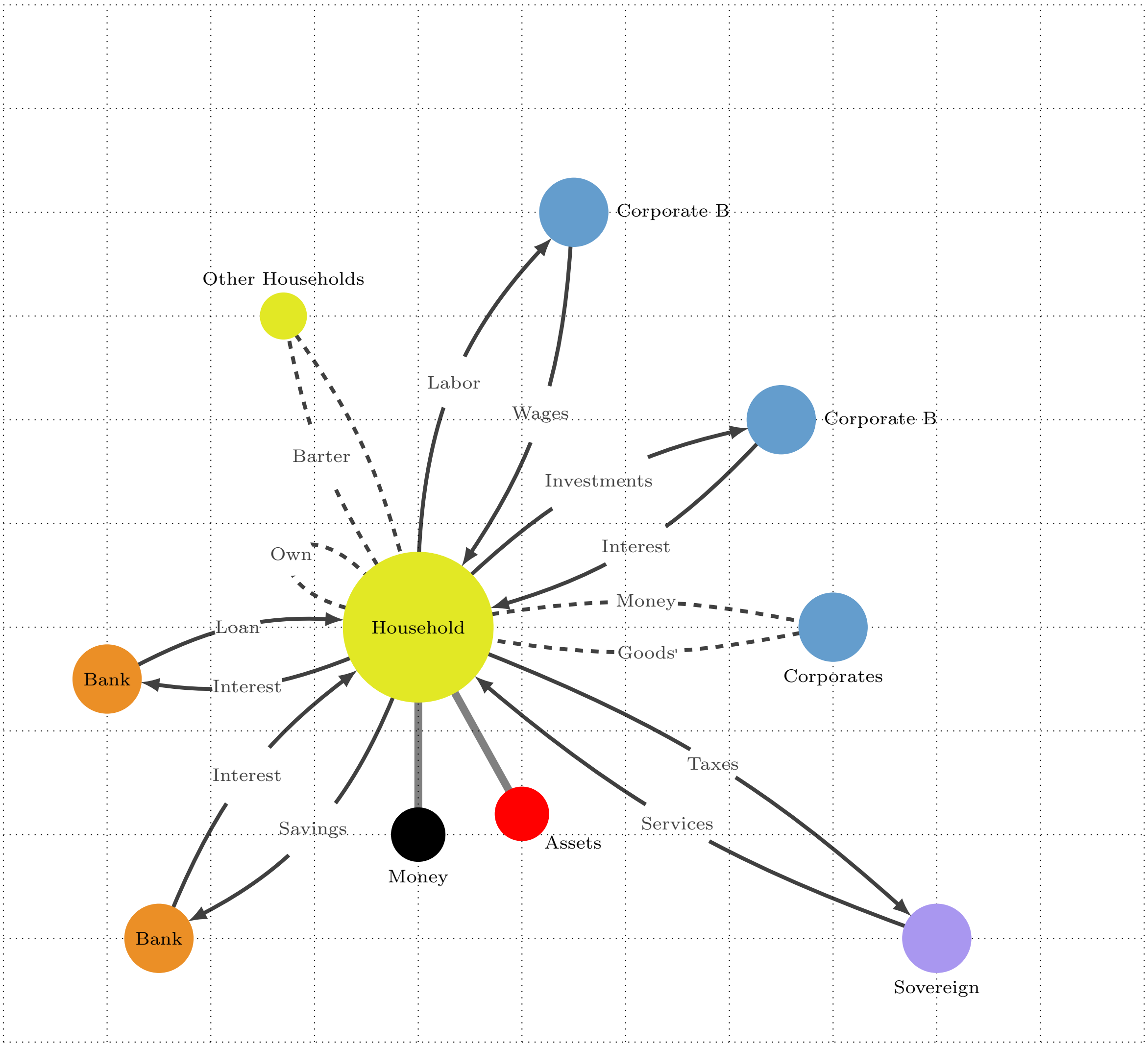
- By allowing nodes and edges of *different types* (e.g. individuals versus companies or companies in different sectors)
- By allowing a variety of qualitative and/or quantitative information to be associated with nodes or edges (e.g. assets, contracts, company statuses etc)
- By capturing the *temporal dimension*, i.e., the evolution of an economic graph in time.

Formally property graphs are *directed, labeled and attributed multi-graphs.* This means roughly the following:

- The *attributed* adjective means that both nodes and edges carry associated information (attributes or properties) that can be significantly more detailed than the simple existence of a node or a relation between nodes.
- The nature of the attributes (e.g. the data type) is in principle quite flexible including both numerical and categorical types.
- The *multi-graph* adjective means that there are multiple possible edges between nodes expressing different types of relationships. The number of possible edges between nodes is not constrained.
- The *labeled* adjective means that both nodes and edges are individually identifiable and may belong to distinct types.

In the visual example below we illustrate the significant additional richness of a property graph. The graph focuses on the *economic neighborhood* of a single household (a family) as an economic entity. A variety of economic (e.g contractual) relations with other agents are depicted as edges of different types (exchange of money or goods, cash flows linked to agreements and legal contracts etc). The graph also illustrates the concept of *node properties* (in this case just two elements: *money* and *assets*). Edges may have arbitrary properties too, capturing the quantitative and qualitative elements of the network relations.

The total network is composed of an number of distinct types. The network may be completely abstract (focusing on economic relations) but it may also include spatial information (e.g. the location of a particular node) that allows mapping the structure onto a geographical map. Finally all nodes and edge properties may have temporal tags (timestamps) that place the network in temporal continuum.



Mathematically a property graph is defined as follows[2]:

- V is a set (collection) of network nodes. It is decomposed as $V = V^1 \cup V^2 \cdots \cup V^p$, where p is the total number of different *node types*.
- $\bar{x}^p = \{x^1, x^2, \ldots, x^{n(p)}\}$ is a set of n *node attributes*. Different node types will in general have different number $n(p)$ of attributes. We will suppress this general notation for simplicity.
- Collecting the node properties for all similar nodes of type p creates a *data frame* $\mathbf{x}^p_n$ that is, a $n(p) \times N(p)$ matrix of values ($n(p)$ columns, $N(p)$ rows) where $N(p)$ is equal to the number of nodes of a given type.
- $E \subseteq V \times V$ is a set of edges connecting different nodes. The set of edges E is decomposed as $E = E^1 \cup E^2 \cdots \cup E^q$, where q is the total number of different *edge types*. Edge types may be node specific (that is specific to a combination of node types) or universal (able to connect all nodes).

[2] This a simplified summary that is adequate for our current purposes. It is worth mentioning that property graphs are both mathematical structures and realizations of concrete *graph databases*. In this white paper we are mostly interested in the structure as a data container

- A set of adjacency matrices A^q (one for each edge type). Adjacency matrices are binary matrices consisting entirely of $(0, 1)$ values. The values A^q_{ij} indicate that a connection of type q exists between nodes i and j but do not provide other qualitative or quantitative information. This role is played by *edge attributes*.
- $\bar{y}^q = \{y^1, y^2, \ldots, y^{m(q)}\}$ is a set of $m(q)$ edge attributes of edges of type q. Different edge type q will in general have different number of attributes $m(q)$.
- The data frame $\mathbf{y}^q_m$ is a $m(q) \times M(q)$ matrix of values ($m(q)$ columns, $M(q)$ rows) where $M(q)$ is equal to the total number of edges (of a given type) that appear in a network.

Property graphs are very general structures including the simpler mathematical graph families as special cases. E.g. a weighted or valued graph is one where the edge properties are assigned a single scalar value.

> We can now state schematically that concentration indexes will (very generally) be maps from the space $(\mathbf{x}^p_n, \mathbf{y}^q_m, A^q)$ of node and edge properties and adjacency structures to real numbers, in short: $I = F(x, y, A)$.

3.1.1 Economic activity representations

The precise nature of node and edge attributes used to represent economic networks is open ended and left to the analyst to design as required. The structure enables abstractions that represent many of the diverse forms of human economic activity, different types of transactions, contracts, accounting approaches etc. In general economic agents might be mapped into nodes but not every node need be an economic agent). Other important entities or concepts may also be usefully considered as a network node - ultimately nodes are book-keeping devices. For example an asset may be considered a node property or a standalone node with its own attributes depending on the fidelity required.

Transactions, contracts or other relations between nodes will be mapped into edges (links). More concretely for our current purposes of integrating the diverse range of concentration indexes we will idealize the following:

1. Individual economic agents, both physical persons and legal only entities will be thought as nodes of the network.
2. Nodes have individual properties, from within a vast variety of types, which are associated with each agent (node). They represent for example *ownership* of assets including cash / money.
3. Agents may engage in *exchanges* (property transfers, service provision etc.) that are abstracted as transactions.
4. Agents may also enter into contracts (that basically govern *future transactions*) and may have finite duration (such as debt) or perpetual nature (such as equity).

We will focus on concentration measurement operations that exclusively rely the above data, that is, we assume that the economic network expressed as property graph contains all the information of interest in the set $(\mathbf{x}^p_n, \mathbf{y}^q_m, A^q)$. It is, though, conceivable and actually quite frequent, that values of interest are not *primary quantities* (directly obtained from measured data) but are derived by processing the above structure. We will see several important examples in the sequel. To avoid overburdening the notation such derived data will be considered part of an expanded graph that *appends* any further derived node or edge attributes to the existing ones.

3.1.2 Contracts

As a concrete and important example we discuss the representation of financial contracts. Contracts are legal constructs that are widely used in modern economies and govern the exchange of various assets and services. Contracts can be seen as collections (bundles) of forward transactions sets. We have already seen transactions as exchanges of cash, real assets, services, other contracts etc. Contracts are term-sheets spelling out a sequence of such future events.[3] There is an enormous variety of contracts and many are relevant in the context of capturing node dependencies in an economic network.

Simplified representations of contracts are possible and may be adequate for various purposes. In the simplest case contract details can be captured as the numerical attributes of edges (e.g. including a list of scheduled cash flows and payment dates), with further logic of the contract encoded implicitly in the *contract type*.[4]

A contract can thus be represented as a relation between two nodes that encodes scheduled future transactions (exchanges) between the two nodes. An important aspect of contracts is that they have duration (maturity), which may in some special cases be infinite (perpetual). The contractual maturity is a future time $t_M > t$ when the final scheduled transaction *must* take place. A contract might entail a transfer from entity i to entity j of an asset a_k (can be a real asset, cash, e-money, services or anything else that is part of the economy) at contract inception $T^t_{ij}(a_k)$. In subsequent times it may stipulate transfer from entity j to entity i (hence reverse) of other assets a_l (a real asset, cash, e-money, services) $T^{t_M}_{ji}(a_l)$. The final scheduled forward transaction specified in a contract determines its *maturity*.

Let us mention that the above conceptual constructs are but a convenient approximation. Besides the caveat of qualitative considerations that may not at all be amenable to quantitative measurement, other economic relations may need more elaborate structures to represent economies and their interactions more faithfully. For example even a simple loan guarantee is a *tri-partite* relation that links a lender, a borrower and a guarantor in a triangle of specific roles. While some aspects of that relation can be modeled using adjacency matrices the intrinsic logic is more complex.

3.2 Network Data Property Classifications

Classic indexes focus on the dispersion of property values along a single, carefully chosen, dimension (which may be associated with a node or link). For example income inequality could be represented as the distribution of an edge property, where the edge represents the salary value associated with a job contract between an individual node and a corporate node. A more comprehensive representation of income might aggregate multiple sources of income represented by other types of edges. To enable using the (in-principle) larger available dataset of an economic network we need to take a step back and explore

[3] We assume here that the entire economic substance of a contact is captured explicitly. In practice such accuracy is only achievable for very restricted sets of monetary exchanges possibly contingent on carefully defined (and legally enforceable) events

[4] A slightly more elaborate specification that still permits implementation in a property graph would involve including *lambda functions* as edge attributes. This is a simple means to encode conditionality (e.g. introduce payments that subject to thresholds and triggers) ([21]).

the nature of available. We turn now on exploring in more detail the collection of node, edge attributes and adjacency matrices $(\mathbf{x}^p_n, \mathbf{y}^q_m, A^q)$ and more specifically their *classification and statistical nature*.[5]

As mentioned already, a fundamental assumption is that any particular economic network expressed as a property graph is but a representative (sample) from a large configuration space. Some measures take this approach literally and associate indexes with *statistical tests* of concentration (or rather lack thereof). We will not in general try to document this view. Another important consideration: Which network elements are considered fixed (constrained) and which are assumed as variable within the permissible *graph configuration space* is very important for normalizing measurement outcomes. We will leave it also out of scope in our survey as it is quite context dependent.

We turn first to classifying node and edge properties as captured by the dataframes $(\mathbf{x}^p_n, \mathbf{y}^q_m)$. In a mathematical sense concentration indexes reflect the structure and properties of the distributions they attempt to summarize. The two important aspects that play a role in this context are:

- The *dimensionality* of the total distribution, namely the shape of the (n, m) tuple of node and edge properties.
- The nature (data type) and range of the random variables making up that distribution, namely whether they are numerical, continuous, categorical or discrete.

We proceed to explore this segmentation in more detail:

3.2.1 Qualitative versus Quantitative Data

A first important distinction is between *qualitative and quantitative data*. This juxtaposition is common in financial / risk management practice. Quantitative variables are generally linked to verifiable *and* numerical measurements (e.g., observed market prices, amounts stated in contracts, numerical figures from audited accounts etc.). Qualitative variables may instead not be directly attributable or linked to concrete empirical observations or the process of their construction is not mechanical / reproducible (e.g. it is the outcome of an expert panel scoring session).

Qualitative does not necessarily mean *subjective* (even more so, less important). Qualitative information such as the text-based description of system properties may be both factual and invaluable for understanding the actual economic network state. Yet such information cannot be used directly in quantitative concentration measurement. *Some* qualitative information may be convertible into the ordinal or categorical variety. E.g. the legal status of a contract with a counterparty may turn from a qualitative variable that is captured e.g., in a paper document into a quantitative state variable in accordance with a predefined categorical scale (performing, defaulted etc.). This conversion process may involve procedures or algorithms of a complexity that excludes it from the core concentration measurement process.

An indicative list of data types that are excluded from further consideration in this paper:

- Any free form text (sentences)
- Strings (words) - unless they are part of a strictly controlled vocabulary (See Categorical Data)
- Binary data (sometimes called blobs or binary large objects)

[5] While the adjacency matrices A^q are obviously crucial for describing complex networks and their concentrations, from a data perspective they are form through pure binary data hence no further classification is required.

- In general any composite objects (e.g, tree structures, that form from combinations of more primitive types

3.2.2 Numerical versus Categorical Data

A most important segmentation for our purposes is the split between *numerical variables* (integers, floating point numbers etc) and *categorical variables* (e.g. a variables that take values from a choice list):

1. **Numerical Variables** are quantitative values of various elementary numerical types that are distributed over a possibly continuous range in one or more dimensions. More complex ranges are definitely possible (half plane / positive only values, some combination of continuous and discrete values etc.) Numerical (including spatio-temporal) properties give rise to empirical distribution functions (one for each node or edge property). Spatial variables are intrinsically two or three dimensional hence the empirical distribution function will typically be higher dimensional.

2. **Discrete Categorical** or Ordinal Variables that take ordered or unordered values from a finite and discrete set. While such variables can use any symbol to enumerate the set elements, by convention we will use integers. Ordinal and categorical variables give rise to categorical distribution functions (the counts of the frequency of occurrence of each possible category value).

Classic concentration measurement tools are generally adapted to the above numerical / categorical split. Some indexes may be applicable to both or be more meaningful for one but not both types of variables. Diversity indexes work in general with categorical variables. On the other hand inequality and concentration risk measurement is in general using numerical variables.

3.2.3 Extensive versus Intensive Variables

In analogy with physical systems, properties of economic networks can be categorized as being either *intensive or extensive*. This characteristic segments variables according to how a quantity changes when the size (or extent) of the system changes. An intensive quantity is one whose magnitude is *independent* of the size of the system. An extensive quantity is one whose magnitude is additive as subsystems are combined. A natural indicator of size is the number of nodes[6] and the intensive or extensive property can thus be expressed in terms of how the variable scales with N.

In general, categorical variables are intensive quantities that do not scale with the size of the system. Some (but by no means all) numerical variables might be extensive. Examples of intensive properties: An interest rate is a numerical property that can be *averaged* across contracts but the *summation* of interest rates is not meaningful per se. Prices are also intensive properties. The sum of prices observed in a market is meaningless but an average price is definitely a useful indicator. Data properties that are *rates or ratios* in nature are likely intensive. Geolocations are further examples of numerical properties that can be averaged (e.g. to identify the clustering of some business activity) but for which the summation or aggregation operation is not informative.

Examples of extensive properties are abundant as well: Any quantities (stocks) of assets owned by economic nodes, notional amounts of contracts associated with edges, cash flow figures involved in transactions between nodes etc. all are examples of extensive properties. Summing up these amounts provides

[6] the number of edges is generally coupled to the number of nodes

a meaningful aggregate that describes the entire system. Extensive properties are positive numbers and can meaningfully be summed across the system. That sum scales linearly with the number of nodes or edges. As mentioned already, an important consideration in the precise form of concentration indexes is to control how such scaling is reflected in the index.

3.2.4 The special case of spatio-temporal variables

Spatiotemporal variables deserve a special mention. Conceptually they can be considered a subset of numerical variables. For example: a timestamp (or a date) is a numerical value conforming to some date format / scheme.

Similarly, spatial (geometric) data may be a tuple of coordinates or a more complex composite of numerical variables expressing, e.g. the latitude and longitude coordinates of the boundaries of a region. The universal and unambiguous meaning of such data within a temporal / spatial context sets them apart from other numerical variables. In practice they also have dedicate set of tools for analysis. To simplify and unify the discussion we will assume that any spatial data properties associated with nodes or edges is also represented as a collection of columns vectors in our set of properties $(\mathbf{x}_n^p, \mathbf{y}_m^q)$.

3.2.5 Stock versus Flow Variables

Economic networks are rather dynamic (changing in time). Every single transaction in an economy is in principle a *graph event*, that is, an occurrence that modifies the graph, even if in a small way. At a more aggregated level as well, dynamics and evolution are common features: Nodes and edges appear and disappear constantly (new generations of individuals, new company formation, new trade relations, adjusted ownership of assets, projects succeeding or failing, maturing contracts etc).

In economic literature such dynamics is conventionally phrased in the language of *stocks and flows*. Stock variables describe the economic *system state* at some point in time whereas flows are the corresponding *changes* to the state (they are the first time derivative). Generally, stock variables at different time instances are connected to their corresponding flows via a difference equation of the following general form[7]

$$S_k^t = S_k^{t-1} + F_k^t \,, \tag{1}$$

where k denotes time instances. The above rather generic equation suggests stocks and flow variables would be typically be governing the evolution of numerical properties and would be of extensive type. Indeed the addition of an incremental amount (flow) to a stock must evidently be meaningful. But temporal different equations will also apply to system-wide (intensive) properties as well. E.g. a central bank policy rate change will get reflected on many contractual rates characterizing distributed edge properties (contracts) across the economic system.

3.3 Old measures in new clothes

We now proceed to list standard considerations from economics and finance and sketch how they might get translated in our graph based computational framework.

[7] We adopt a discrete time framework

3.3.1 Example: Income and Wealth Inequality

Income and wealth inequality are considered in the context of *economic inequality* which can be defined / studied at various levels of aggregation. Here we illustrate how the corresponding quantitative analysis can be rephrased in an economic network language in a hypothetical quantification exercise that uses detailed network data:

- Nodes of type $P = V^1$ represent individuals (physical persons) as economic agents
- A specific node property g capturing owned wealth (the value of real assets)
- Nodes $C = V^2$ representing corporate nodes
- A node property c_j representing total owned assets by a corporate node j
- Edge relations captured by an adjacency matrix A^1 representing *labor contracts* between individuals and corporate nodes
- An edge property I_{ij} that captures income as the form e.g. of monthly or annual salary remitted from corporate entity i to individual person j.
- Edge relations A^2_2 representing *shareholding contracts* between individuals and corporate nodes
- An edge property S that represents the share of corporate node assets owned by an individual

With the above setup, the computation of an inequality index based on income and/or wealth is reduced to aggregating node and edge properties across the network and examining the resulting value distribution. For example the wealth g_i of node i could be computed as the following aggregation[8]

$$g_i = V(\sum_{j=1}^{C} A^1_{ij} I_{ij}) + \sum_{j=1}^{C} A^2_{ij} S_{ij} c_j \tag{2}$$

where C is the number of corporate nodes, and V is a capitalization function applied to the salary figures.

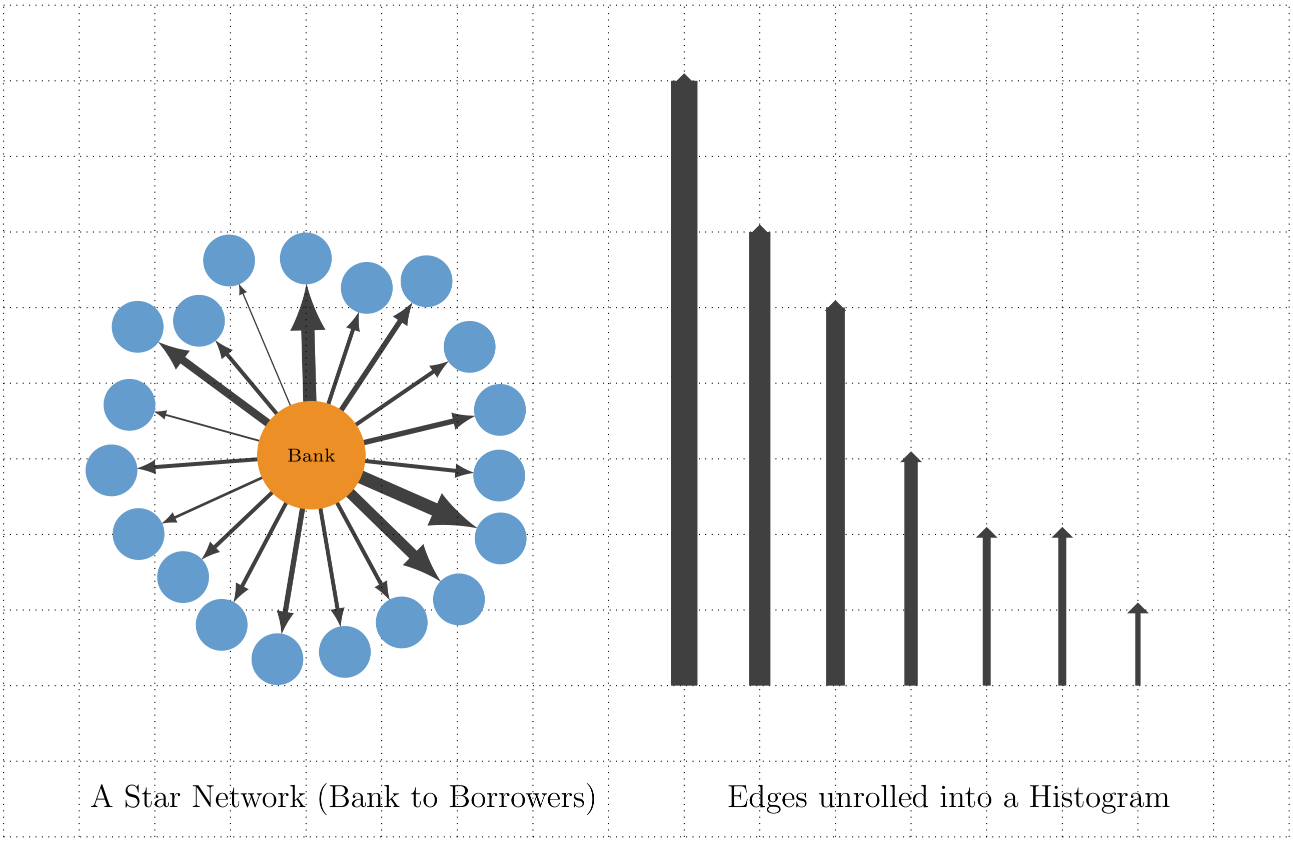
3.3.2 Example: Credit Risk "Name" Concentrations

Classic concentration risk analyses like name sector, geographical or financial product concentration have been traditionally based on computing concentration indexes.

Name concentration is one of the well known types of credit risk concentration. It rests on analyzing various aspects of the distribution of *exposures* (contractual amounts subject to default risk) to individual companies ("names". The classic approach will use as a starting point a table of exposures and an associated graph to visualize the distribution, identify the top exposures or compute a suitable index. How would the same be represented in a "network data" approach?

In a network context a portfolio concentration is represented as a simple *star graph*. The focus is on a central node that is a portfolio holding node (e.g. a bank) entering into contracts with clients. Those contracts are represented as edges, emanating from the bank node and reaching the client node. The classic concentration of a loan book is thus obtained by applying an index to the distribution of an *edge attribute* (for example an attribute that could be labeled "exposure" or "contract nominal value").

[8] Ignoring corporate node liabilities

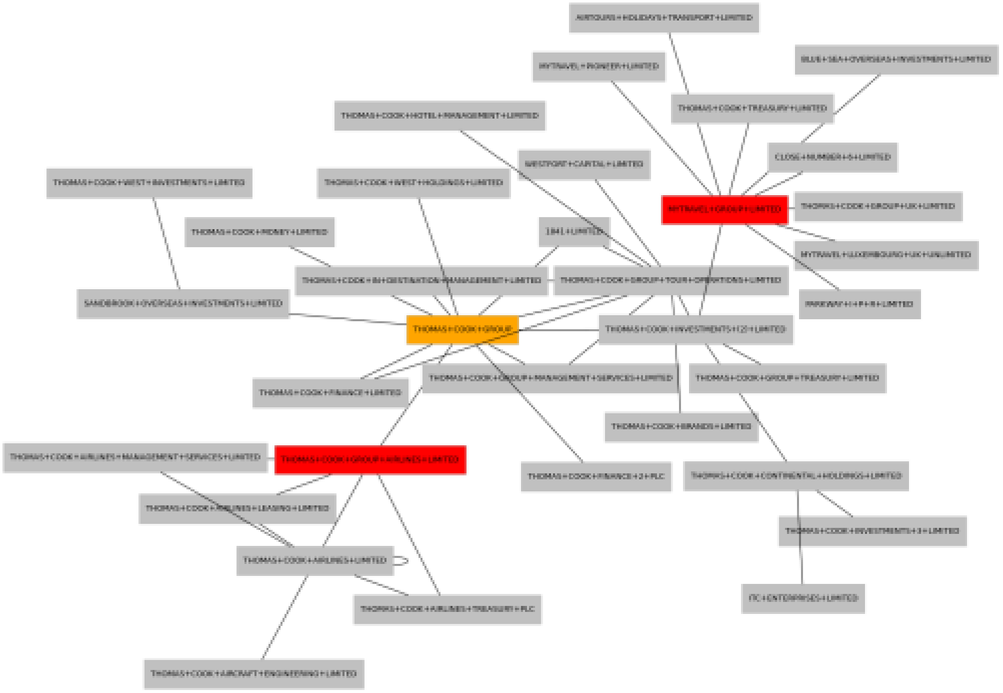


A Star Network (Bank to Borrowers)

Edges unrolled into a Histogram

Visual illustration of a classic concentration analysis that is based on a portfolio picture. At the center of the graph is a bank node and the network is formed by its financial (credit contracts to a collection of other legal entities depicted as a cloud around the central node. Further relationships of these nodes with each other (or other economic agents) are not captured in this picture.

The width of the arrows pointing to borrower nodes is meant to indicate the nominal value of the loan contract. Hence in this case name concentration is the prevalence of credit contracts with large nominal amounts with particular nodes in the network. It is clear that in this case the adjacency matrix is not particularly informative. Yet focusing on what we actually by a corporate node and "exposure" to a counterparty immediately suggests that important aspects of network structure is never far from the surface.


AIRTOURS+HOLIDAYS+TRANSPORT+LIMITED
MYTRAVEL+PIONEER+LIMITED
BLUE+SEA+OVERSEAS+INVESTMENTS+LIMITED
THOMAS+COOK+HOTEL+MANAGEMENT+LIMITED
THOMAS+COOK+TREASURY+LIMITED
WESTPORT+CAPITAL+LIMITED
CLOSE+NUMBER+5+LIMITED
THOMAS+COOK+WEST+INVESTMENTS+LIMITED
THOMAS+COOK+WEST+HOLDINGS+LIMITED
MYTRAVEL+GROUP+LIMITED
THOMAS+COOK+GROUP+UK+LIMITED
THOMAS+COOK+MONEY+LIMITED
1841+LIMITED
MYTRAVEL+LUXEMBOURG+UK+UNLIMITED
THOMAS+COOK+IN+DESTINATION+MANAGEMENT+LIMITED
THOMAS+COOK+GROUP+TOUR+OPERATIONS+LIMITED
SANDBROOK+OVERSEAS+INVESTMENTS+LIMITED
PARKWAY+I+P+R+LIMITED
THOMAS+COOK+GROUP
THOMAS+COOK+INVESTMENTS+(2)+LIMITED
THOMAS+COOK+FINANCE+LIMITED
THOMAS+COOK+GROUP+MANAGEMENT+SERVICES+LIMITED
THOMAS+COOK+GROUP+TREASURY+LIMITED
THOMAS+COOK+BRANDS+LIMITED
THOMAS+COOK+AIRLINES+MANAGEMENT+SERVICES+LIMITED
THOMAS+COOK+GROUP+AIRLINES+LIMITED
THOMAS+COOK+FINANCE+2+PLC
THOMAS+COOK+CONTINENTAL+HOLDINGS+LIMITED
THOMAS+COOK+AIRLINES+LEASING+LIMITED
THOMAS+COOK+INVESTMENTS+3+LIMITED
THOMAS+COOK+AIRLINES+LIMITED
ITC+ENTERPRISES+LIMITED
THOMAS+COOK+AIRLINES+TREASURY+PLC
THOMAS+COOK+AIRCRAFT+ENGINEERING+LIMITED

A network graph illustrating the corporate structure of a major corporate group (Thomas Cook) which filed for bankruptcy in 2019. The different boxes are all the legal entities (companies or special purpose vehicles) that are economically closely related to the business group. The central (orange) node might be what one would term the principal counterparty node. Around it, the graph illustrates a large number of closely associated corporate entities that are linked with various explicit or implicit contractual and economic relations. Aggregating the total or *one obligor exposure* to such a cluster or entities can be quite an intricate exercise.

3.3.3 Example: Geographical Risk Concentrations

Geographical concentration aims to capture any propensity of exposure to entities that operate economically within a particular territory / jurisdiction. Traditional analysis might drastically simplify the picture by assigning nodes to predefined country or regional categories. In an economic network picture geography enters as spatial data characterizing node activities.

Business sector concentration is another classic analysis that can substantially improved in an economic network context. Sector concentration focuses on identifying corporate nodes of a particular industrial type or business model. In turn the risk of such node sub-categories is natively analysed in their own economic network context (e.g., the collection of suppliers, employees and customers).

Finally, financial product concentration is the prevalence of contracts (hence edge relations) with particular features (e.g. certain types of optionality leading to prepayment risk or similar dependence on reference rates leading to joint payment shocks in the case of large interest rate movement).

3.3.4 Example: Market Share Concentration

Market concentration concerns the relative *market share* of a firm. Depending on the context it might be e.g., based on sales figures or the total assets of the firm. In a standard approach one might collect *already aggregated* sales data as reported e.g. in the statement of income account of the corporate entity.

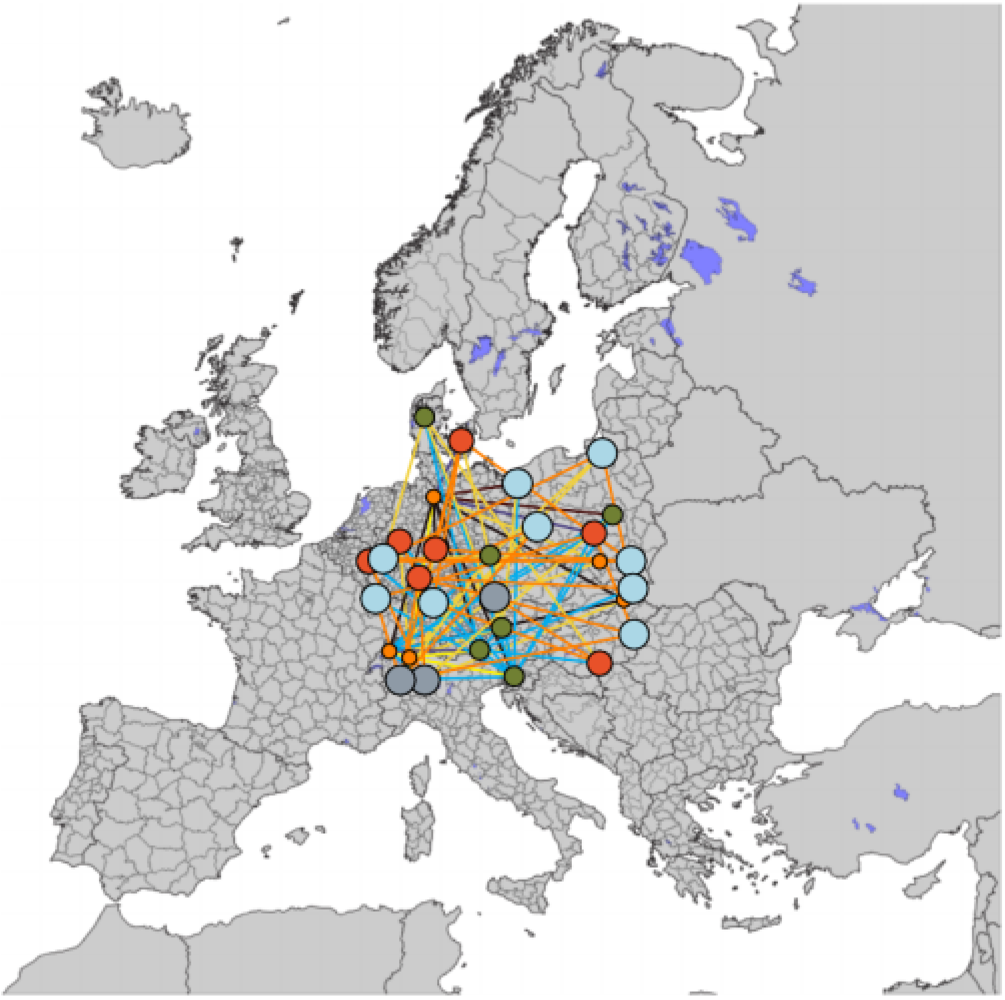


Figure 2: A hypothetical economic network illustrating a number of different data dimensions and how those can be used to map concentration. Spatial coordinates help locate the nodes on a map. Node type and properties can be indicated as node size and color. Edge type and properties are indicated by edge color. Obviously nodes and edges can have many more associated data points than can be legibly visualized. Spatial concentration indexes can in turn be derived on the basis of the distribution of nodes on the map

How can we derive market share from a network representation? This is an example where a procedure might be in-principle possible but not available in practice due to data limitations.

Let us assume that the business model involves manufacturing and sales of gadgets. To represent sales transactions within the graph framework we image these as instant actions expressed as edge relations between buyer and seller. A snapshot graph view will in general not be representative of the network structure. The additional element required to achieve a better measure is the *temporal aggregation* of network data. Schematically the aggregate sales S_i over a period for entity i can be expressed as :

$$S_i = \sum_{k=1}^{T} \sum_{j=1}^{N} A^k_{ij} y^k_{ij} \tag{3}$$

where k is an index of the number of temporal snapshots where A^k_{ij} is the adjacency matrix linking seller i and buyer j and y^k_{ij} is the monetary value of that sale transaction.

3.4 Is it a node or an edge concentration?

As discussed in [18], contractual relationships that are naturally seen as edge properties may actually get mapped into node properties in the context of *balance sheet accounting*. In a balance sheet representation the edge relations of the reporting node get aggregated into node properties. We saw already two examples of this transformation: the balance sheet representation of a bank node converting contracts from edges to balance sheet items and the income statement of a corporate converting sales transactions into aggregated income streams.

4 Index Construction Mechanics

As mentioned indexes are maps from the space $(\mathbf{x}^p_n, \mathbf{y}^q_m, A^q)$ to the real numbers (schematically the value of the index is $I = F(x, y, A)$). The choice of the functional form F is essentially free (which is not particularly helpful). Useful suggestions generally constrain this functional space through various specifications. The following are three of the most important categories of considerations

- identifying the set of readily available input data. While data collection has been famously much expanded with the advance of digitization, detailed, low-level network data are still the exception rather than the rule

- specifying a *tractable set of operations* on the network data. Tractability is also a moving target and context dependent. It may refer to computational constraints such as performance or algorithmic complexity but also to issues such as transparency and explainability

- specifying a set of desirable axioms / requirements that the index must satisfy. This will typically be driven by the domain where the index is primarily used

4.1 Data Objects

Let us summarize and visualize the discussion so far before we embark on discussing the mechanics of constructing indexes.

4.1.1 Node and Edge Dataframes

Node and edge properties data are represented as a collection of tables. For example:

Node	Float	Integer	Categorical	Timestamp	Latitude	Longitude
1	123.3	10	2	1623577438	31.9	-4.8
2	4.21	1	0	2743821235	41.2	2.6
...	34.1	3	1	2123527438	11.4	6.2
N	10.2	15	2	1623577438	24.5	40.1

The above table collecting a set of network node data illustrates the distribution of numerical, categorical and spatio-temporal properties for a certain type of node. All nodes of same type have (in principle) the same number of attributes[9]. After standardization, all datasets have a representation in terms of *digits* (including any categorical and temporal data).

An edge collection will have in principle a very similar data frame as shown below. The main differences are that i) the data frame row index associated with and edge will range over $M = N^2$ (the product of relevant node counts) and ii) data properties that make sense for edges might be quite different from those of nodes. For example spatial data associated with nodes might be spatial points. Instead, an edge might future an array of data points, expressing e.g., *a way or route relation* between nodes.

Edge	Float	Integer	Categorical	Timestamp	Start	End
1	123.3	10	2	1623577438	(31.9, 1.2)	(-4.8, 10.2)
2	4.21	1	0	2743821235	(41.2, 4.5)	(2.6, 0.9)
...	34.1	3	1	2123527438	(11.4, 5.4)	(6.2, 20.1)
M	10.2	15	2	1623577438	(24.5, 3.5)	(40.1, 1.2)

4.1.2 Adjacency Data

The third set of data objects after node and edge dataframes are the adjacency matrices representing relations of various types. In standard graphs the relevant object is a single *adjacency matrix*, a square matrix A^{ij} with elements either zero or one, representing whether nodes (i, j) are connected.

In most realistic economic networks one would have to introduce multiple edge types to track relevant connections. This can be thought of as a set of matrices, or an *adjacency tensor* (A^{ij}_k). The adjacency tensor captures which *type of edge* exists between which pairs of nodes. Further complexity is introduced if permissible edge types are contingent on node types (e.g. only bank nodes can provide saving accounts to individuals) but there is no need in the current context to further burden the notation.

The following table illustrates an adjacency tensor. Links between nodes are indicated by 1. The diagonal elements are zero as there are no ”self-loops”. Undirected graphs where the direction of the relationship is not relevant or meaningful can be represented via symmetric matrices. A loan relationship might be better modeled as a directed edge (distinguishing lender / borrower).

[9] It is conceivable that some will have null values or missing data

Edge Type-1	1	2	...	N
1	0	0	1	0
2	1	0	1	0
...	1	1	0	1
N	1	0	0	0
Edge Type-2	1	2	...	N
1	0	1	0	0
2	0	0	1	1
...	1	1	0	1
N	0	1	0	1

4.2 Data Operations

We now have an overview of the economic network data structure and how it represents potentially interesting concentrations. Yet constructing concrete measures starting with the above raw data may still involve potentially significant intermediate operations. We classify such operations as follows:

- Simple collection (aggregation) of values from relevant nodes / edges
- Univariate operations on collected values (sorting, binning, calculation of proportions / weights)
- Multivariate operations on collected values (combining data to construct e.g. a spatial weights matrix)
- Local adjacency matrix operations in the network neighborhood (see example of collecting wealth contributions)
- Global operations in the network (as required for example to identify the importance of nodes according to some centrality measures)

We will next describe in some detail operations as those are required for the indexes we will catalog in the last section.

4.2.1 Categorical Distributions and Counts

A categorical variable C_i associated with the i-th node or edge ranges over some menu of options that can be *encoded* as integer values $(1, \dots, S)$. In the case of ordinal variables this set of integers is assumed to also have a concrete order (we will not make use of that sub-class here). Concentration analysis of categorical properties is based on the *fraction of occurrence* (abundance) of variable realizations over node or edge elements. This type of categorical variable distribution is very common in general and largely the default option in biodiversity studies. But categorical data are also common in economic / financial context. As an example, the type of collateral used in a loan contract is a categorical variable that characterizes an edge (relation) between two nodes (lender and borrower).

For categorical properties the count of occurrences N_r of the r-th attribute value are used to create the distribution vector:

$$c_r = \frac{N_r}{N} \tag{4}$$

where

$$N_r = \sum_{i=1}^{N} \mathbb{1}_{\{C_i = r\}} \tag{5}$$

and

$$N = \sum_{r=1}^{S} N_r \tag{6}$$

is the total number of nodes or edges under consideration

NB: Once we have converted categorical properties into weights c_r, the machinery of indexes is identical with that of numerical variables that we will see below and many index formulas are in principle the same.

4.2.2 Numerical Distributions

In several classic inequality / concentration indexes the core mathematical object is the distribution of a numerical variable X taking N ”values” $x_i, i \in [1, N]$ over a population of nodes or edges. Here the main requirement to compute an index is for the data element to admit a meaningful *summation* to a total figure:

$$x_T = \sum_{i=1}^{N} x_i \tag{7}$$

The summation operation allows the definition of *weights* for the i-th node or edge numerical property:

$$w_i = \frac{x_i}{x_T} \tag{8}$$

which then forms the basis of the concentration index computed from the vector w.

Numerical and Categorical approaches to concentration indexes are related given that any numerical property can be converted into an ordinal property by *binning* or coarse-classifying the variable range and classifying any realization into the corresponding discrete values. The binning process defines thresholds $[H_r, H_{r+1}]$ where $r \in [1, S]$ for the numerical variable x using some specified algorithm (which may or may not depend on the dataset itself) and then assign individual observations x_i to the appropriate bin r. Counting the occurrences C_r within each bin creates the required categorical (actually ordinal) variable as per above.

There are two important caveats: The process of binning *loses some information* which may or may-not be acceptable for the use case and the binning is an additional modeling element (assumption) that may interfere (e.g. introduce bias) in the concentration analysis.

The table below illustrates the similarities and difference of categorical versus numerical variables.

Variable Type	Numerical	Categorical
Value	x_i	N_r
Index Range	$i \in [1, N]$	$r \in [1, S]$
Sum	$x_T = \sum_{i=1}^{N} x_i$	$N = \sum_{r=1}^{S} N_r$
Population Size	N	N
Categories	-	S
Weight	$w_i = x_i / x_T$	$c_r = N_r / N$
Average Weight	$\mu = X_T / N$	$\mu = 1/\text{S}$

4.2.3 Categorical Clustering of Spatial Data

> *"Step back and ask, what is the most striking feature of the geography of economic activity? The short answer is surely concentration". Krugman (1991).*

Multivariate considerations are quite common in the context of spatial analysis. Multivariate datasets can always be treated using univariate (marginal) methods that explore concentration along one dimension at a time. Take for example the spatial distribution of a specialized economic activity: the surface area of commercial real estate units within a region. The data set might involve three data vectors: surface area per location and the x, y coordinates of the location. Three marginal views are possible: the surface area distribution (irrespective of location) and the location distribution along either the x or y dimensions. Yet the complete picture clearly is only revealed once all three dimensions are considered. First, using the x and y coordinates simultaneously will reveal whether there is true spatial clustering. Using all three data series may also reveal, for example, that location clustering is correlated (or anti-correlated) with surface area clustering.

A widely used class of multi-variate measures captures sectoral / geographical concentration. More concretely, a distribution fo numerical values (for example a value attached to nodes representing companies and expressing company size in terms of FTE) alongside a business sector and the geographic association of each company.

In order to evaluate the geographic distribution of establishments economists have historically first employed cluster-based methods, i.e., they measure the spatial concentration of economic activity according to pre-defined geographic limits (regions, countries etc).[10] In a multivariate approach data might be grouped along categories. Let us use a typical three-dimensional set:

Node (i)	Measurement	Sector (s)	Geography (a)
1	10.1	1	1
2	7.2	1	1
3	3.4	2	1
4	2.7	2	2
5	1.1	2	2
...	...	...	...
N	21.1	S	G

Two integer map functions $S(i), A(i)$ classify each node into a respective sector and geography. This arrangement allows the computation of multiple indexes using the categorical or numerical avenue we already discussed. In a clustering approach, each node property (measurement) is associated with one industry (business sector) and one geography. We have N measurements E_i (one value per node, for definiteness let us assume it is a measure of "size", e.g. number of employees, turnover etc) which can be aggregated as follows:

[10] This leads to the well known Modifiable Areal Unit Problem (MAUP) which can be summarized as sensitivity to the shape, size, and position of the geographical units used

$$E_T = \sum_{i=1}^{N} E_i \tag{9}$$

$$E^{s\bullet} = \sum_{i=1}^{N} E_i \, \mathbb{1}_{\{S(i)=s\}} \tag{10}$$

$$E^{\bullet a} = \sum_{i=1}^{N} E_i \, \mathbb{1}_{\{A(i)=a\}} \tag{11}$$

$$E^{sa} = \sum_{i=1}^{N} E_i \, \mathbb{1}_{\{S(i)=s,A(i)=a\}} \tag{12}$$

In the above, the total measured size is E_T, the aggregate size of each industry across all areas is $E^{s\bullet}$ where the bullet denotes an implied summation over all areas. The total size per area is $E^{\bullet a}$. Within each area, an industry comprises of N^{sa} exposures, summing up to a total of E^{sa} for each industry / area combination. From the absolute values aggregated above we can derive various fractional values[11]

$$w_i = \frac{E_i}{E_T} \tag{13}$$

$$z_i = \frac{E_i}{E^{S(i)\bullet}} \tag{14}$$

$$q_i = \frac{E_i}{E^{\bullet A(i)}} \tag{15}$$

$$y^s = \frac{E^{s\bullet}}{E_T} \tag{16}$$

$$x^a = \frac{E^{\bullet a}}{E_T} \tag{17}$$

$$h^{sa} = \frac{E^{sa}}{E^{s\bullet}} \tag{18}$$

In the above, w_i is the usual proportion of a node across the entire network as already discussed for numerical variables. The fractional size of a node within its sector is z_i and within its area it is q_i. Further, x^a is the fraction of each area as part of the total and y^s the fraction of each sector as part of the total. The allocation to both industry and areas is given by h^{sa}.

As an example, given the above proportions one can construct a number of different HHI type metrics:

$$H = \sum_{i=1}^{N} w_i^2 \tag{19}$$

$$H^s = \sum_{i=1}^{N} z_i^2 \, \mathbb{1}_{\{S(i)=s\}}, \; s \in [1, S] \tag{20}$$

$$H^a = \sum_{i=1}^{N} q_i^2 \, \mathbb{1}_{\{A(i)=a\}}, \; a \in [1, A] \tag{21}$$

$$H^G = \sum_{a=1}^{A} (x^a)^2 \tag{22}$$

$$H^S = \sum_{s=1}^{S} (y^s)^2 \tag{23}$$

[11] As before, had the measurement been categorical in nature, we would have to work directly with proportions

The classic HHI index capturing concentration across the entire node collection is denoted as H. HHI indexes capturing node concentration within each industry (respectively area) are defined as H^s and H^a. The geographic concentration of nodes (irrespective of industry) is captured by H^G. The HHI index capturing concentration in the distribution of different industry sectors is H^S.

We see here both the power and weakness of simple approaches when in the presence of multi-dimensional data: marginal (one dimensional) approaches are easy to implement providing immediate insights, yet there is a proliferation of measures with potentially conflicting messages.

4.2.4 Spatial Weights Matrix

A spatial weights matrix denoted as $w_{ij}(r)$ is a derived data structure that is seen in operations involving spatial data. In the simplest case it is equal to 1 if a suitably defined distance between the two entities i and j is less than a radius r (0 otherwise).

$$w_{ij}(r) = \begin{cases} 1 & \text{for } d(i,j) \geq r \\ 0 & \text{for } d(i,j) < r \end{cases} \tag{24}$$

where $d(i,j)$ could be for example the Euclidean distance that is constructed using pairs of numerical values representing coordinates:

$$d(i,j) = \sqrt{((x_1^i - x_1^j)^2 + (x_2^i - x_2^j)^2)} \tag{25}$$

A spatial weights matrix need not be based on distances but any measure of spatial association. For example, if a node i is located within a polygon P (territory), the spatial weights matrix linking to another node j within a polygon Q may be determined by whether P and Q are contiguous territories (have a common border). The spatial weights matrix is a special case of a parametric adjacency matrix. It is not defined from economic relations as the example we have seen so far but is inferred from input spatial data.

4.2.5 Computing centrality measures

Last but not least an important class of operations on adjacency matrix data. The structure (topology) of the adjacency matrix is a core subject of *network science*. Many approaches to network concentration analysis compute a *centrality measure* per node and subsequently explore the distribution of that measure (thus again decoupling the problem and reducing it to a univariate analysis).

A most fundamental tool is counting the number of edges emanating from a node. This is termed the node degree $d_j = \sum_{i=1}^{N} A_{ij}$. The presence of nodes with different degrees (the distribution of d_j over the network) means that nodes may have very different *roles* within the economic network. This is described as different *network topologies.* In network studies this and other *centrality measures* are designed to help rank / classify network nodes based on their topological importance. There is a wide variety of such measures:

Box 2. Centrality measures are non-trivial examples of derived graph data that reflect the propensity and distribution network links

In network theory, indicators of *centrality* assign numbers or rankings to nodes reflecting their role within a network. There are many examples:

- Degree centrality: Defined as the number of links incident upon a node
- Closeness centrality: The average length of the shortest path between a node and all other nodes in the graph
- Betweenness centrality: The number of times a node acts as a bridge along the shortest path between two other nodes
- Eigenvector centrality: Derived from global influence characteristics of different nodes

All these examples can be abstracted as a function from adjacency data to a numerical range $d_i = \phi_i(A)$. For example the number of counterparties of a financial intermediary would be simply the degree centrality computed on edges representing credit contracts.

5 Index Catalog

The index catalog is a list of commonly used metrics across a variety of domains, cast in a uniform notation and with reference to the network data structures already discussed in the previous sections.

The list is not exhaustive but aims to cover the major families. Within each discipline there are variations in particular in scaling conventions that may subtly modify the circumstances for which each expression is suitable. Therefore there is no attempt to reduce the collection to ”genuinely different” approaches.

5.1 Index Functions

The general (symbolic) form of a network concentration index computed on the basis of graph property data has been mentioned already:

$$I^{pq} = F(\mathbf{x}^p_n, \mathbf{y}^q_m, A^q) \tag{26}$$

where $(\mathbf{x}^p_n, \mathbf{y}^q_m)$ are sets (dataframes) of property data vectors populated with either numerical or categorical values, A is a set of adjacency matrices with binary values and pq filters the set of nodes and edge connection by type. This form appears sufficiently general but in practice it is constrained by the specification of F. In particular complex pre-processing operations on network data that concern e.g., multiple node or edge types or involve complex algorithms would in practice better handled outside the index definition.

In more pedantic representation the index function is recast as:

$$I = F(w_1, \ldots, w_n, c_1, \ldots, c_m, A_1, \ldots, A_q) \tag{27}$$

where w_n are vectors of numerical variable weights, c_n are vectors of categorical (encoded as integers) proportions and A_q are matrices of boolean variables. In practice most commonly used indexes are fairly simple functions focusing on slices of the available dataset but we already saw multi-dimensional examples.

5.2 Index Function Sub-Categories

Popular indexes can be classified according to various attributes. The following will be *use oriented* segmentation:

- General Purpose Indexes using on any one of the w, c data vectors. This category forms the majority of common indexes. In this use case, some property of the system is studied in isolation from other properties and/or the network topology. The most developed theory and practice of concentration indexes concerns the one-dimensional case but some applications require higher dimensional considerations. The vast majority of *classic* concentration measures are univariate risk measures. They are a map from a one-dimensional distribution that captures some property of interest into a real number (typically positive). Nevertheless there are some important bivariate examples in use and spatial concentration indexes can be two or three dimensional.
- Diversity Indexes are distinct as the use exclusively categorical c vector data. While their form is quite close to general purpose indexes, they are special in that both the population size N and the number of categories S may be relevant and used in the construction of the index.
- Temporal Clustering which is a special univariate numerical analysis.
- Spatial Concentration. Multivariate Concentration Indexes. Two or more properties from the (w, c) set are studied jointly. This segment includes Spatial Concentration Indexes which capture the density of objects (or object properties) in two or three-dimensional space.
- Network Concentration Indexes that are calculated using the adjacency matrices A and (optionally additional information from the numerical or categorical vectors X, Y).

5.3 General Purpose Indexes

General purpose indexes focus on a single variable (univariate, one-dimensional) and can be used for either numerical or categorical data.

5.3.1 L-Zero Norm

The l^0 norm is a traditional sparsity measure. It is the count of non-zero elements in a vector. This is obviously meaningful for numerical data only.

$$l^0(w) = \#\{w_j \neq 0, j = 1, \dots N\} \tag{28}$$

5.3.2 Range

The range is defined as the absolute difference between the highest and lowest weights divided by the mean proportion. The notation has been discussed already in 4.2:

$$R = \frac{\max(w) - \min(w)}{\mu} \tag{29}$$

5.3.3 Variance

The standard statistical variance serves also as a basic measure of concentration or dispersion.

$$V = \frac{1}{N}\sum_{i=1}^{N}(w_i - \mu)^2 \tag{30}$$

5.3.4 Coefficient of Variation

The coefficient of variation, also termed *relative standard deviation*, is derived from the variance.

$$CV \equiv \frac{\sqrt{V}}{\mu} \tag{31}$$

5.3.5 Standard Deviation of Logarithms

For numerical values spanning a large range taking the logarithm may be appropriate:

$$H = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(w_i) - \log(\mu))^2} \tag{32}$$

5.3.6 Berger-Parker Index

The Berger-Parker index is simply the maximum w_i or c_r value in the dataset. Hence it is the largest weight or abundance (frequency) of an attribute of either a node or edge in the network.

$$\mathrm{BP} \equiv \mathrm{CR}_1 \tag{33}$$

This index is used also in its inverse form.

5.3.7 Concentration Ratio

For numerical attributes of nodes or edges, the definition of the **concentration ratio of order** a is the sum of the a-th largest weights (assuming a sorted set of observations):

$$\mathrm{CR}_a = \sum_{i=1}^{a} w_i \tag{34}$$

The index can be applied to either node or edge properties and it is very popular for stylized values e.g. $a = 1, 4, 10, 20$ at it provides an intuitive and immediately understood measure of concentration.

Two points worth mentioning here: the ordering (sorting) of the vector w and the presence of an index parameter a.

See Also Concentration Ratio

5.3.8 Relative Mean Deviation

The relative mean deviation (or *Pietra Coefficient*)[22]

$$RMD = \frac{\sum_{i=1}^{N}|w_i - \mu|}{2|\mu|} \tag{35}$$

It is equal to the maximum distance between the Lorenz curve and the equal distribution line.

5.3.9 Mean Log Deviation

The mean log deviation (also *Thiels L* [23]) is given by:

$$L = \frac{1}{N} \sum_{i=1}^{N} \log(\frac{\mu}{w_i}) \tag{36}$$

5.3.10 Herfindahl-Hirschman Index

The widely used Herfindahl-Hirschman Index is expressed as

$$\mathrm{HHI} = \sum_{i=1}^{N} w_i^2 \tag{37}$$

For categorical attributes one replaces w_i with the corresponding category count c_r.

See Also HHI

5.3.11 Simpson Index

The Simpson Index is related to the HHI

$$\mathrm{D}_1 = 1 - \sum_{i=1}^{N} w_i^2 \tag{38}$$

$$\mathrm{D}_1 \equiv 1 - \mathrm{HHI} \tag{39}$$

It is also used as the inverse Simpson index:

$$\mathrm{D}_2 = \frac{1}{\sum_{i=1}^{N} w_i^2} \tag{40}$$

The above examples show that indexes are used in variety of very closely related forms.

5.3.12 Hall-Tideman Index

The HTI index is defined as

$$\mathrm{HTI} = \frac{1}{2\sum_{i=1}^{N} i w_i - 1} \tag{41}$$

An alternative name is the *Rosenbluth Index*.

5.3.13 Gini Index

The Gini index is defined as

$$G = \frac{1}{N} \sum_{i=1}^{N} (1 - 2i)\, w_i + 1 \tag{42}$$

It is worth mentioning that the Gini index has been reformulated in dozens of different expressions [24].

See Also Gini

5.3.14 Kolm Index

Another important parametric class is that of the Kolm indexes [25]

$$I_\alpha = \frac{1}{\alpha} \log \left(\frac{1}{N} \sum_{i=1}^{N} e^{\alpha[w_i - \mu]} \right) \tag{43}$$

where is α is a parameter that may be assigned any positive value.

5.3.15 Hannah-Kay Index

The (generalized) Hannah-Kay index is defined as [26]

$$\mathrm{HK}_a = \begin{cases} (\sum_{i=1}^{N} w_i^a)^{1/(1-a)} & \text{for } 0 \leq a \neq 1 \\ e^{(\sum_{i=1}^{N} w_i \log w_i)} & \text{for } a = 1, \end{cases} \tag{44}$$

We can think of this index as a generalization of the HHI index (which is a special case for $a = 2$). The reciprocal HK index is also used.

See Also Hannah-Kay

5.3.16 Tsallis Entropy

The Tsallis entropy of order a is defined as

$$H_a = \frac{1}{a-1}(1 - \sum_{i=1}^{N} w_i^a) \tag{45}$$

This is used also as the Hill-Tsallis index

$$N_a = (1 - (a-1)H_a)^{1/1-a} \tag{46}$$

5.3.17 Atkinson Index

The parametric Atkinson index is given by [27]

$$A_a = \begin{cases} 1 - N^{a/(a-1)} \left(\sum_{i=1}^{N} w_i^{1-a} \right)^{1/(1-a)} & \text{for } 0 \leq a \neq 1 \\ 1 - N\, e^{\left(\frac{1}{N} \sum_{i=1}^{N} \log w_i \right)} & \text{for } a = 1 \end{cases} \tag{47}$$

5.3.18 Gaussian Entropy

The Gaussian entropy is defined as [28]

$$H_G = \sum_{i=1}^{N} \log w_i^2 \tag{48}$$

5.3.19 Renyi Entropy

The Renyi entropy is defined as

$$H_a = \frac{1}{1-a} \log(\sum_{i=1}^{N} w_i^a) \tag{49}$$

Used also as the Hill-Renyi index

$$N_a = e^{H_a} \tag{50}$$

5.3.20 Shannon Index

The Shannon Index is the Shannon entropy. Alternative names: Shannon-Wiener.

$$H = -\sum_{i=1}^{N} w_i \log w_i \tag{51}$$

See Also Shannon

5.3.21 Theil Index

The Theil T index is the Shannon index with a sign reversal

$$\mathrm{T} \equiv -\mathrm{H} \tag{52}$$

See Also Theil

5.3.22 Hoyer Sparseness

The Hoyer Sparseness is defined as [29]:

$$\mathrm{HS} = \frac{\sqrt{N} - \frac{\sum_i^N |w_i|}{\sqrt{(\sum_i^N w_i^2)}}}{\sqrt{N} - 1} \tag{53}$$

5.3.23 L-p Norm

Many of the named indexes are based on some expression involving the l^p norm of the data vector:

$$\|w\| = (\sum_i^N |w_i|^p)^{1/p} \tag{54}$$

5.3.24 Generalized Entropy Measures

Measures related to the concept of entropy can be integrated into a generalized entropy expression:

$$\mathrm{GE}_a = \begin{cases} \frac{1}{Na(a-1)} \sum_{i=1}^N ((Nw_i)^a - 1), & a \neq 0, 1 \\ \sum_{i=1}^N w_i \log(Nw_i), & \alpha = 1 \\ -\frac{1}{N} \sum_{i=1}^N \log(Nw_i), & \alpha = 0 \end{cases} \tag{55}$$

- For $\alpha = 0$ it becomes the mean log deviation.

- For $\alpha = 1$ it becomes the Thiel index or Shannon entropy
- For $\alpha = 2$ it becomes the half-squared coefficient of variation.

5.4 Diversity Indexes

5.4.1 Margalef Diversity Index

For a categorical variable that has S category types distributed over nodes or edges and N total measurements the Margalef index is simply:

$$D = \frac{S-1}{\ln N} \tag{56}$$

5.4.2 Menhinick Diversity Index

The Menhinick Diversity metric comes also from biodiversity studies and is rather similar to the Margalef index:

$$D = \frac{S}{\sqrt{N}} \tag{57}$$

5.4.3 McIntosh Index

The McIntosh Index for categorical variables is [30]:

$$\mathrm{D}_{Mc} = \frac{1 - \sqrt{\sum_{r=1}^{S} c_r^2}}{1 - 1/\sqrt{N}} \tag{58}$$

NB: This is similar to the HHI index but applied to category abundances but adjusts for sample size N.

5.4.4 Pielou Evenness Index

Pielou's measure of "species evenness" divides the Shannon index by the natural logarithm of the number of categories S

$$H = -\sum_{r=1}^{S} c_r \log c_r \tag{59}$$

$$J = \frac{H}{\ln(S)} \tag{60}$$

5.4.5 Brillouin Index

The Brillouin Index is nearly identical to the Shannon-Wiener entropy but is based directly on category counts (not proportions) and explicitly accounts for

$$H_B = \frac{\log(N!) - \sum_{i=1}^{N} \log(N_r!)}{N} \tag{61}$$

5.5 Temporal Clustering

Temporal clustering is a special type of univariate distribution where the variable is a date, timestamp or other indicator of time.

5.5.1 Tango Temporal Clustering Index

Tango proposed the following quadratic form as an index for the level of clustering in time [31]. All observations are grouped into m equal temporal intervals. The frequency count is $N_1, \ldots, N_m$. The weight vector is $c_m = N_m/N$ where N is the total number of observations. The index is given by

$$T = \sum_{k=1}^{m} \sum_{l=1}^{m} c_l D_{lm} c_m \tag{62}$$

where D_{lm} is a (temporal) distance matrix. A standard expression is $D_{lm} = |l - m|$, the index difference between temporal intervals.

5.5.2 Greenwood Statistic

The Greenwood statistic is a *spacing statistic* that can be used to evaluate clustering of events in time (also linear clustering in space). In general, for a given sequence of events in time or space the statistic is given by

$$\mathrm{GS} = \sum_{i=1}^{N} w_i^2 \tag{63}$$

where w_i represents the interval between events or points in space and is a number between 0 and 1 such that the sum of all $w_i = 1$. The formal appearance of the formula is identical to the HHI index but the use context focuses on distributional properties under the assumption of independent arrivals.

5.6 Spatial Concentration

Spatial concentration is generally a multi-dimensional concept. It involves at least a value measurement and a location. In the most common use case of spatial economics the typical dataset is three-dimensional. Historically the analysis of spatial concentrations has developed three major approaches (in order of increasing complexity):

- Purely clustering (or categorical) methods where spatial data is aggregated implicitly in ”regions”. This procedure converts the geospatial locations of measured values into categorical variables.
- *Spatial Weight Matrix* methods, where continuous distance measurements are converted into spatial adjacency matrices
- Continuous methods that do not truncate spatial data but use them directly, e.g. kernel function approaches that estimate distribution densities.

We only include in the catalog the first two types as continuous methods add a significant computational layer.

5.6.1 Ellison-Glaeser Index

The Ellison-Glaeser Index (EG) is an index developed for the assessment of industrial agglomeration [32]. Using the notation of 4.2.3, the EG industrial concentration index (per industry s) is given by

$$\gamma^s = \frac{G^s - (1 - H^G)H^s}{(1 - H^G)(1 - H^s)} \tag{64}$$

where G^s is a metric capturing industry concentration per area

$$G^s = \sum_{a=1}^{A} (h^{sa} - x^a)^2 = \sum_{a=1}^{A} (\frac{E^{sa}}{E^{s\bullet}} - \frac{E^{\bullet a}}{E_T})^2 \tag{65}$$

5.6.2 Maurel and Sedillot Index

The Maurel-Sedillot Index for a sector k is a variation on the EG index given by [33]:

$$\gamma_k = \frac{G_k - H}{1 - H} \tag{66}$$

where G is a measure of geographic concentration:

$$G_k = \frac{\sum_r^M s_{kr}^2 - \sum_r x_r^2}{1 - \sum_r^M x_r^2} \tag{67}$$

and

$$H = \sum_j w_i^2 \tag{68}$$

is the standard HHI index.

5.6.3 Getis and Ord G Statistic

The Getis and Ord G statistic measures the degree of association that results from the concentration of weighted points within a radius r from the point i. It is defined as [34]:

$$G_i(r) = \frac{\sum_j w_{ij}(r)\, x_j}{\sum_j x_j} \tag{69}$$

where $w_{ij}(r)$ is a symmetric binary spatial weight matrix with ones for all links defined as being within distance r of a given node i and all other links are zero including the link of node i to itself. The numerator is the sum of all x_j within r of i but not including x_i. The denominator is the sum of all x_j not including x_i.

A global G across the network is given by:

$$G(r) = \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij}(r) x_j}{\sum_{i=1}^N \sum_{j=1}^N x_i x_j}, \quad i \neq j \tag{70}$$

5.6.4 Moran's I

Moran's I is defined as

$$I = \frac{N}{W} \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2} \tag{71}$$

where w_{ij} is again a suitably defined spatial weight matrix and $W = \sum_{i=1}^N \sum_{j=1}^N w_{ij}$.

5.6.5 Geary's C

Geary's C is defined as

$$C = \frac{(N-1) \sum_{i=1}^N \sum_{j=1}^N w_{ij}(x_i - x_j)^2}{2W \sum_{i=1}^N (x_i - \bar{x})^2} \tag{72}$$

5.6.6 Marcon and Puech M-Functions

The Marcon and Muech M-Functions allow the evaluation of the relative geographic concentration and co-location of industries in a non-homogeneous spatial framework. They are defined as [35]

$$M_S(r) = \frac{\sum_{i=1}^{N_S} \frac{\sum_{j=1,j\neq i}^{N_S} w_{ij}(r)x_j}{\sum_{j=1,j\neq i}^{N} w_{ij}(r)x_j}}{\sum_{i=1}^{N_S} \frac{X_S - x_i}{X - x_i}} \tag{73}$$

Where N nodes (companies, projects), $w_ij(r)$ the spatial weights matrix, N_S nodes of sector type S, x_i the numerical attribute of entity i, X_S the numerical value of sector S and X the aggregate numerical value.

5.7 Adjacency Clustering Measures

In the last sub-category of measures in our catalog we focus on pure *adjacency based measures* which are generically of the form $I = F(A^q)$. To start with, any of the centrality measures mentioned in 4.2.5 can be analyzed for dispersion / concentration using the general indexes of section 5.3. In the sequel we describe some more specialized measures.

5.7.1 Graph Density

Graph density (or Edge density) is a metric that comes from network theory and aims to capture the prevalence of connectivity within a graph. It is defined as:

$$D = \frac{M}{N(N-1)} \tag{74}$$

where $N = \dim(A)$ is the number of nodes (of some type p) and $M = \sum_i \sum_j A_{ij}$ is the number of edges (again of some type q) in G. High graph density means economic agents have rich inter-dependencies on each other. This ratio indicates how close the graph is to a complete graph (a complete graph has edge density 1).

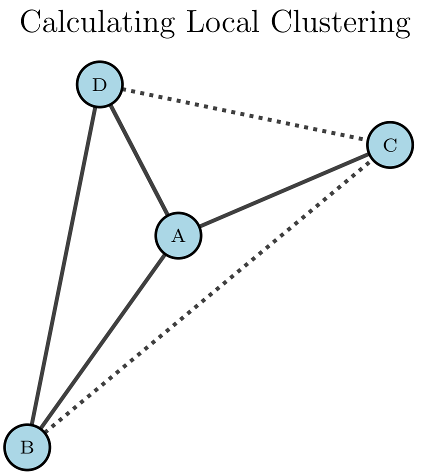
5.7.2 Global Clustering Coefficient

For each node i, the local clustering coefficient, CL_i is the fraction of pairs of neighbors of i that are connected within than set. The maximum number of possible links between the neighbors of node i is simply $d_i(d_i - 1)/2$ (where d_i is the degree - number of edges emanating from node i).

The (local per node) clustering coefficient expresses the fraction of actual connections over the maximum value. Let $v_n \subseteq V$ be the set of nodes connected to i-th node. That is, $v_j \in v_n$ when $A_{ij} = 1$. Then $c = \sum_{jk} A_{jk}$ where the pair of (j, k) ranges over all nodes in the neighborhood v_n is the number of closed triangle paths.

$$\mathrm{CL}_i = \frac{2c}{d_i(d_i - 1)} \tag{75}$$

In the below example with four nodes (A,B,C,D) the local clustering coefficient of node A is 1/3 (the two dotted edges indicating the additional two potential connections that would bring the clustering coefficient to unity)



It is also expressed as:

$$\mathrm{CL}_i = \frac{\sum_{j,k} A_{ij} A_{jk} A_{ki}}{\sum_i d_i(d_i - 1)}, \quad j,k \neq i \tag{76}$$

where the numerator counts the number of triangles in which node i participates. The global clustering coefficient is then given by

$$\mathrm{CL} = \frac{\sum_{i,j,k} A_{ij} A_{jk} A_{ki}}{\sum_i d_i(d_i - 1)}, \quad j,k \neq i \tag{77}$$

The average clustering coefficient is simply the average of CL_i over all nodes:

$$\mathrm{CL}_A = \frac{1}{N} \sum_{i=1}^{N} \mathrm{CL}_i \tag{78}$$

5.7.3 Network Entropy

An entropy like measure based on the adjacency matrix can be computed by assigning a random walk probability from node to node on the basis of its degree d_i. More specifically if we define transition probabilities as per

$$p_{ij} = \begin{cases} 0, & A_{ij} = 0, \\ 1/d_i, & A_{ij} = 1 \end{cases} \tag{79}$$

then a network entropy is defined as [36, 37]:

$$H = \frac{1}{N \ln(N-1)} \sum_i^N \ln(d_i) \tag{80}$$

References

[1] M.D. Konig and S. Battiston. From Graph Theory to Models of Economic Networks. A Tutorial. *A.K. Naimzada et al. (eds.), Networks, Topology and Dynamics*, 2009.

[2] M.D. Flood. On Counterparty Networks. 2015.

[3] G. Marti et al. A review of two decades of correlations, hierarchies, networks and clustering in financial markets. *Preprint*, 2017.

[4] P. Csermely et al. Structure and dynamics of core/periphery networks. *Journal of Complex Networks*, 2013.

[5] C. Joyez. Network Structure of French Multinational Firms. *Preprint*, 2017.

[6] Z. Poszar et al. Shadow Banking. *FRBNY Staff Reports*, 2012.

[7] J. Duesenberry. A Process Approach to Flow-of-Funds Analysis. *National Bureau of Economic Research*, 1962.

[8] G.L Breton L.B Duc. Flow-of-funds analysis at the ECB. Framework and Applications. 2009.

[9] Financial Stability Board. The Financial Crisis and Information Gaps. *Report to the G-20 Finance Ministers and Central Bank Governors*, 2009.

[10] A.G. Haldane. On microscopes and telescopes. *Bank of England Speech*, 2015.

[11] D.D. Gatti H. Dawid. Agent-Based Macroeconomics. *Universitat Bielefeld: Working Papers in Economics and Management*, 2018.

[12] J.P. Gleeson T.R. Hurd. A framework for analyzing contagion in banking networks. *Preprint*, 2011.

[13] E. Cerutti and H.Zhou. The Global Banking Network in the Aftermath of the Crisis. *IMF Working Paper*, 2017.

[14] S. Boccaletti et al. The structure and dynamics of multilayer networks. *Physics Reports*, 2014.

[15] M. Kivela et al. Multilayer networks. *Journal of Complex Networks*, 2014.

[16] H. Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30:348–361, 1920.

[17] P. Papadopoulos. Open Risk WP1: Revisiting Simple Concentration Indexes . 2015. Online Link.

[18] P. Papadopoulos. WP8: Connecting the Dots: Economic Networks as Property Graphs. *Open Risk White Papers*, 2019. Online Link.

[19] A. Ghrab et al. GRAD: On Graph Database Modeling. *Preprint*, 2010.

[20] R. Angles. The Property Graph Database Model. *Preprint*, 2018.

[21] Open Risk Blog. Open Source Securitisation, 2019. Link.

[22] G. Pietra. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, page 775, 1914.

[23] H. Thiel. *Economics and Information Theory*. North Holland, Amsterdam, 1967.

[24] P. Verme L. Ceriani. Individual diversity and the gini decomposition. *World Bank Policy Research Working Paper*, 6763, 2014.

[25] S-C. Kolm. Unequal inequalities. *Journal of Economic Theory*, 12:416–442, 1976.

[26] J.A. Kay L. Hannah. *Concentration in Modern Industry: Theory, Measurement and the UK Experience*. Mac Millan Press, London, 1977.

[27] A. Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2:244–263, 1970.

[28] K. Kreutz-Delgado and B.D. Rao. Measures and algorithms for best basis selection.

[29] P.O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[30] R.P. McIntosh. An index of diversity and the relation of certain concepts to diversity. *Ecology*, 48:392–404, 1967.

[31] T. Tango. The detection of disease clustering in time. *Biometrics*, 40:15–26, 1984.

[32] L. Glaeser G. Ellison. Geographic concentration in u.s. manufacturing industries: A dartboard approach. *Journal of Political Economy*, 105, 1997.

[33] B. Sedillot F. Maurel. A measure of the geographic concentration in french manufacturing industries. *Regional Science and Urban Economics*, 29:575–604, 1999.

[34] J.K. Ord A. Getis. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24, 1992.

[35] F. Puech E. Marcon. Measures of the geographic concentration of industries: Improving distance-based methods. *Journal of Economic Geography*, 10:745–762, 2010.

[36] G.S. Freitas et al. A detailed characterization of complex networks using information theory. *Nature Scientific Reports*, 2018.

[37] M. Small. Complex networks from time series: Capturing dynamics. In *2013 IEEE International Symposium on Circuits and Systems*, 2013.