How to attribute credit if you must

by Luke S Meiklejohn

MKLLUK001

lukemeiklejohn@gmail.com

https://lukemeiklejohn.github.io

Submitted to the University of Cape Town

In partial fulfilment of the requirements for the degree

Master of Philosophy in Financial Technology

African Institute of Financial Markets and Risk Management, Faculty of Commerce

UNIVERSITY OF CAPE TOWN

Date of Submission: January 20, 2020

Supervisor: Assoc. Prof. Co-Pierre Georg

DECLARATION

I, Luke Meiklejohn, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: .....Signed by candidate.....

Date: 20 January 2020

Abstract

Data ownership is of fundamental importance in the digital economy of today. Commercializing academic research, whilst maintaining ownership of it, is a task that can now be accomplished due to the strengths of blockchain technology, which allows data to be registered, made unique, and traced to its origins.

We propose a blockchain use-case for licencing academic research, based off an academic project named UniCoin.

In this thesis, we discuss how to fairly attribute credit between all sources of knowledge that contribute to new pieces of academic research, using citation network analysis and centrality measures. Katz centrality, in-degree centrality, and PageRank are three potentially useful centrality measures, with varying results: these are compared using case studies based on three papers co-authored by Andrei Shleifer. We use these centrality measures to guide how to fairly attribute credit, and thus how to distribute licencing revenues generated through UniCoin.

Keywords: credit diffusion, citation networks, blockchain, research marketplace.

Acknowledgements

To my family, and in particular my parents, Julie and Shaun Meiklejohn, for always putting my education first, and believing that I would excel even when I doubted myself.

To Mo Swainston, for being a critical part of my support system, and seeing me through the completion of not one but two challenging degrees.

To Chris Maree, for the invaluable programming skills you have taught me throughout the year, and for inspiring and challenging me to further myself. It has been a huge pleasure working with you, and this thesis would not be what it is without you.

To my supervisor, Associate Professor Co-Pierre Georg, for pushing me out of my comfort zone: this project was challenging, in a completely new subject matter to me, but I am proud of the finished product.

To Adam Butler, for taking the time to provide thoughtful feedback on my final drafts.

And to my employers at the First Rand Group for funding this degree.

I could not have gotten this far without the support.

Contents

1 Introduction

This thesis discusses the allocation of credit between academic papers, aided by the construction of a citation network and a potential blockchain-based credit allocation system. The thesis is born from an academic project called UniCoin[1], a decentralized smart contract-based non-custodial research marketplace which allows academics to benefit from commercially viable research. The aims of UniCoin are a) to allow researchers to sell licences to intellectual property, and b) ensure that any funds received are distributed fairly. In this paper we extend on this project by further investigating how to fairly attribute credit between all those who contribute to a paper, to guide how to distribute licencing revenue. This research is done in parallel to research by Maree (2020), who investigates the former aim of UniCoin.

There are at least two sources of knowledge that result in the creation of a published work – the authors of that work, and all authors cited within it. Researchers customarily cite all published works they use while writing a paper. All academics who significantly contribute to the paper are listed as authors. It is common for papers to be authored by multiple academics, and for papers to include many cited works as references – as noted by Fang (2018), references serve as a useful way to help convince readers that a paper's argument is sound, and including multiple reputable references helps a paper do this well.

The literature has done significant work with citation and authorship networks. Kim and Diesner (2014), and Tol (2011) discuss the allocation of credit between co-authors – how to attribute credit between those who authored it, in a fair way which benefits those who contributed the most. The issue of diffusion of credit in a citation network, acknowledging the papers that contributed to another paper, has also been discussed (see Radicchi et al. (2009) and Wang et al. (2016)). Significantly

[1] http://unicoin.win

less work, however, has been done into a combination of the two – how to allocate credit fairly between all contributors, acknowledging the value added by a co-author as well as the existing knowledge provided by a cited work. Fang (2018) proposes a framework to do this, which has not yet been implemented empirically, but will be discussed. Additionally, we discuss how credit can be given to third parties who would not previously acknowledged, such as lab assistants.

We use centrality measures in citation networks to guide how to allocate credit between papers cited by a paper, and discuss how to allocate credit between cited papers and the authors of the original paper. We discuss how a blockchain-based platform could be developed to enable researchers to licence their work while retaining ownership of it, and ensure that licencing revenues are distributed fairly. To test the framework empirically, we use three papers authored (in part) by Andrei Shleifer to visualise the results of this allocation.

2 Academic norms and measuring impact

In academic writing, it is the norm to recognize prior work upon which one's writing depends, through citing that work. Gilbert (1977) argues that the award of a citation is more than recognizing the property rights of other academics, but is a function of persuading readers that one's argument is valid. Academic writing that furthers existing research can trust that the prior research is true, and use it as the foundation of a new argument, demonstrating the novelty of the new research and how the authors are contributing to the academic landscape. Or, authors can use citations to justify their positions or statements, strengthening their argument through substantiation from a trusted source (Gilbert, 1977).

The 'trusted' nature of the source is critical to ensuring that the resulting argument is accepted as valid, and citing respectable sources adds more value than irrelevant or untrustworthy sources. Indeed, one function of a peer-reviewed journal is to establish the trustworthiness of the research published therein.

Moreover, the awarding of a citation to a published work can also be seen as a self-fulfilling prophecy, in that by awarding a paper a citation, one is recognizing that it is a trusted and esteemed source, but also *making* it more trusted and esteemed. In this sense, the number of citations a paper or academic has received is seen as a proxy for the scientific impact that they have. Many papers have been written attempting to quantify the impact of an academic or a paper through the number of citations they have received - introducing measures such as the significant h-index due to Hirsch (2005), the g-index due to Egghe (2006), and more recently the Euclidian index due to Perry and Reny (2016)[2] and the A-index due to Stallings et al. (2013), amongst others.

[2]Notably, the inspiration for the title of this thesis.

These indices aim to summarize a researcher's academic track record with some numerical score. For instance, the h-index tells us that a researcher with a score of h has at least h publications which have received at least h citations (Hirsch, 2005), whereas the g-index tells us that a researcher with a score of g has at least g publications with a sum of at least g^2 citations between them (Egghe, 2006). The varying definitions of each index can result in varying rankings when applied to a set of researchers, as demonstrated by each of the papers above introducing the indices, but they all have in common that they use citations as an input for measuring impact.

Additionally, centrality measures have also been applied to citation networks. In network theory, centrality measures relate a given node to their role in the entire network (Jackson, 2010) - one simple example being degree centrality, which measures how well a given node is connected to other nodes. Other measures, such as the Bonacich network centrality measure, use the structure of the citation network to determine who the most important players are - it turns out these are the players who are connected to the most important players (West et al., 2010). In a citation network, this makes sense - if a paper is cited by many important papers, then it is surely also an important paper. Similarly to the Bonacich measure, both Google's PageRank algorithm due to Page et al. (1999) and the Eigenfactor metric due to West et al. (2010) use the concept of eigenvector centrality to determine impact. Where the Eigenfactor metric was initially developed to rank journals, providing an index online at http://eigenfactor.org, it has been adapted to an article level.

Index-based measures of impact, due to their definitions, discard a lot of information. For instance, we defined the h-index above - it only considers h of an academic's n potential publications, and has many shortcomings. An example that illustrates this is a researcher with a few papers that are cited to a much higher degree than

the rest of their papers - suppose they have 20 papers with 20 citations each, but 5 with 200 citations. Their h-index will be limited to 20, significantly lower than it perhaps should be (Stallings et al., 2013). Young researchers, with very few papers, will also be disadvantaged - if someone has only two papers, but each has received many citations, their h-index will be only two.

For this reason, and due to the fact that these indices are commonly applied to researchers rather than on a paper-level, we rather look closer at centrality measures in this paper.

2.1 Centrality as a proxy for impact

There are many measures of centrality, each of which captures slightly different information, and as such, the use of such a measure should be justified by the context. Degree centrality, mentioned above, captures only how many nodes a node is connected to, but not the importance of those nodes. Closeness centrality measures how close a node is to other nodes in the network, and betweenness centrality measures how central a node is in terms of connecting other nodes (Jackson, 2010). All three of these measures are fairly simple to calculate and can give good insights into the structure of a network.

As we have already mentioned, in a citation network, the importance of a paper should be determined by the importance of the papers which cite it, and this simple motivation leads us to look more closely at the following few centrality measures.

Citation networks have some interesting properties. Where co-authorship networks represent authors as nodes, we represent papers as nodes. Firstly, these networks are dynamic, can only increase in size over time (Portenoy et al., 2017), and have directed edges. An edge illustrating that paper A cites paper B means it is (almost)

impossible[3] for paper B to cite paper A - these networks are acyclic. An edge is represented by $g_{ij} \in \{0, 1\}$ where

$$g_{ij} = \begin{cases} 1 & \text{if } j \text{ cites } i \\ 0 & \text{otherwise} \end{cases}$$

An adjacency matrix $\mathbf{G}$ has g_{ij} in the (i, j)-th position. In an undirected network, this would be a symmetric matrix, and g_{ij} would be 1 simply if there was an edge between the two nodes. $\mathbf{G}^k$ is the k-th power of $\mathbf{G}$ and consists of elements $g_{ij}^{[k]}$, the number of walks of length k from i to j.

Consider the example of Figure 1 below.

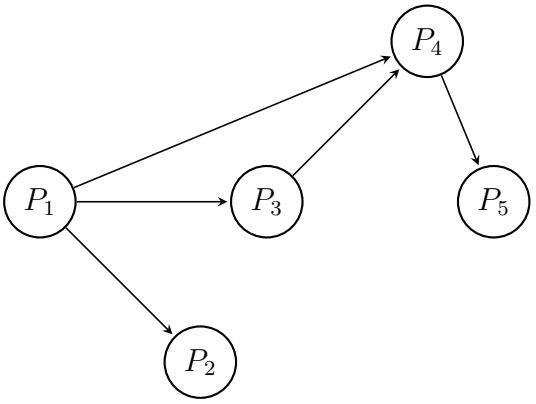

Figure 1: An example citation network.

Here, P_1 cites P_2, P_3, and P_4. P_3 cites P_4, and P_4 cites P_5. The adjacency matrix is:

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

[3]Whilst not technically impossible, it is very rare for this to occur, unless two papers are published in a collaboration.

There are many ways to analyse such a network to determine the most important nodes. One very simple centrality measure is in-degree centrality: the number of in-edges (citations), normalized by the maximum possible degree. Where $d_i^*(g)$ is the number of in-edges of paper g, the in-degree centrality is thus $d_i^*(g)/(n-1)$ in a network with n nodes.

Whilst the number of edges linked to a node may well be a good indication of the importance of that node, a few measures take this concept further, and say that what is more important is the *type* of nodes which are linking to a node.

The Katz prestige draws on the idea that a node's centrality, or importance, should be influenced by the importance of the nodes it is connected to, and the Katz prestige of a node is defined at the sum of its neighbours' Katz prestiges, divided by their relevant degrees (Rusinowska et al., 2011). That is,

$$P_i^K(g) = \sum_{j \neq i} g_{ij} \frac{P_j^K(g)}{d_j(g)}$$

Such a measure makes sense in the application of citation networks: a paper's prestige is influenced by the prestige of its citing papers, but to a lesser degree by a given paper if that paper cites many papers. The above can be rewritten as

$$P^K(g) = \hat{g}P^K(g) \Rightarrow (\mathbf{I} - \hat{g})P^K(g) = 0$$

Where $\mathbf{I}$ is the identity matrix and $\hat{g}$ is the matrix of adjacencies scaled by their degrees, that is $\hat{g}_{ij} = \frac{g_{ij}}{d_j(g)}$. So, solving for the Katz prestiges of a network corresponds to solving for the unit eigenvector of $\hat{g}$.

Bonacich extended on this, proposing that the centrality of a paper should be proportional to the centrality of the nodes it is connected to (Jackson, 2010), through finding the eigenvector of the unscaled matrix g. Where $C^e(g)$ is the eigenvector

centrality of g, we write

$$\lambda C_i^e(g) = \sum_j g_{ij} C_j^e(g)$$

In matrix notation we can write $\lambda C^e(g) = gC^e(g)$. That is, $C^e(g)$ is an eigenvector of g, and λ is its eigenvalue (Jackson, 2010) - generally the largest eigenvector is taken. So, Katz prestige is a form of eigenvector centrality where the network adjacency matrix has been adjusted.

A third formulation can also be considered, where the prestige of a node is a function of the walks that emanate from it - weighted in terms of the length of the walk. That is, we define the matrix

$$\mathbf{M}(g, a) = \sum_{k=0}^{\infty} a^k \mathbf{G}^k$$

Here, a is some small decay factor, and this reduces the contributing weight of longer paths. So, the elements of this matrix are $m_{ij}(g, a) = \sum_{k=0}^{\infty} a^k g_{ij}^{[k]}$, a weighted sum of the paths from node i to j.

The Bonacich centrality of node i is then defined as $b_i(g, a) = \sum_{j=1}^{n} m_{ij}(g, a)$, that is, the total number of (weighted) paths that emanate from i (Ballester et al., 2006). This is sometimes referred to as Katz-Bonacich centrality, as in Bloch et al. (2016), and sometimes as Katz Centrality, as in Newman (2010). To maintain consistency with the software we later use, we refer to it in this paper as Katz Centrality.

In matrix notation, this can be written

$$b(g, a) = [\mathbf{I} - a\mathbf{G}]^{-1} \cdot \mathbf{1}$$

Katz centrality discounts paths in proportion to their length. In citation networks, this makes sense: in Figure 1, $P3$ should pass more influence to $P4$ than it passes to $P5$.

Katz prestige and Katz centrality are both variants of eigenvector centrality. One further variant that can be considered in citation networks is PageRank due to Page et al. (1999). PageRank was famously the originating concept behind Google, and drives how web pages are ordered in search results. Just as a website can be viewed as important if it has many 'backlinks' (that is, in-edges - pages that link to that page), an academic paper is important if it has many citations. PageRank also incorporates the notion that an in-edge from an important node should be more important than many in-edges from obscure nodes.

Where N_u is the number of links from paper u, and B_u is the set of pages that link to paper u, the simplified version of PageRank for a paper u is defined as

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

Here, c is a normalization factor to ensure "the total rank of all webpages is constant" (Page et al., 1999).

The full PageRank algorithm models the behaviour of a random surfer. An internet user follows a walk of links between webpages, randomly clicking on one on each page. However, with some probability they will break out of this cycle and choose a new random page to visit. The probability that a user continues in the walk is referred to as the damping factor. Ultimately, the final ranking given to a node will correspond to the probability "that a random walk will be at that node after a sufficiently long time" (Page et al., 1999).

So, we have discussed a few centrality measures - ways to determine which are the most 'important' nodes in a network, and in our context, which papers should receive a higher amount of credit in a flow of credit from another paper. Once that credit is allocated to a paper, the next decision is how to allocate credit amongst its authors.

2.2 Accounting for co-authorship

From the perspective of the cited paper, receiving many citations is a sign of impact, whereas from the perspective of the citing paper, the awarding of a citation is to pay homage to the intellectual heritage of the citing paper and recognize the impact of the cited paper (Kostoff, 1998). On a more local scale, authorship is awarded to recognize the role played by academics in the writing of a paper.

Across disciplines, collaboration is increasingly common, and this poses a problem when attempting to fairly distribute credit in a citation network. If a portion of credit is allocated to a paper, it still needs to be determined how best to allocate this amongst the paper's co-authors.

A few potential solutions exist. The first is the egalitarian option: to share credit equally, claiming it isn't possible to determine the proportions in which each author contributed. This approach does introduce problems and can benefit secondary authors to the detriment of primary authors (Kim and Diesner, 2014).

Alternative counting methods have been introduced so as to counteract this, which allocate a decreasing amount of credit to authors as their position in the ordering of authors decreases (rank weights). These approaches assume that authors are listed in decreasing order of their contributions, and as Tol (2011) notes, are *ad hoc* and unsatisfactory, because approaches to listing authors vary by discipline - noting that in economics it is common to list authors alphabetically. It is also difficult to capture the true contribution with a formula that decreases in a predefined manner.

Tol (2011) introduces a new method using 'Pareto weights', where the amount of credit allocated to an author relates to the probability of that author attaining the number of citations that the paper in question has attained. Suppose a widely-cited Professor and a less-cited younger author collaborate on a paper. If that paper

receives many citations, most of the credit flows to the Professor, whilst if the paper receives relatively few citations, most of the credit will flow to the younger author.

This property could be either an advantage or a disadvantage, but since it relates the question of attributing credit not to the share of contribution of each author, but to the response to the paper by the academic community, its validity is not clear-cut. They note that egalitarian weights are a reasonable approximation to Pareto weights, and for the purposes of this paper we will assume egalitarian weights as the method of sharing credit between co-authors.

3 A blockchain solution

The motivation for this paper stemmed from the project UniCoin, as previously mentioned, which functions as a decentralized marketplace for licences to use research for commercial purposes.

UniCoin aims to provide researchers with ownership of their research, and the ability to grant licences to corporate entities wishing to use their work for commercial purposes. Where Maree (2020) looks at the pricing of data in a digital marketplace, here we briefly evaluate the motivation for a blockchain-based platform to do this.

3.1 An introduction to blockchain

A blockchain is an immutable append-only public ledger of transactions which have been agreed upon through a consensus mechanism by the majority of the participants in a system. A key property of a blockchain is that it is a decentralized system, as opposed to centralized (Drescher, 2017). In a centralized system, players in a network are connected through one central player, who has some element of control or coordination over the network, whereas in a decentralized system, no one individual player holds such an element of control.

In a blockchain, the players who maintain the network are referred to as nodes. All nodes maintain their own copy of the ledger. These nodes control which transactions are approved, and subsequently recorded in the ledger. These transactions are stored in 'blocks', which are appended to prior blocks, leading to a growing 'chain' of records (Zheng et al., 2017). An illustration of this process from Nakamoto (2008), whose white paper proposed Bitcoin, is shown in Figure 2. Nodes use a protocol to make sure that all nodes agree upon which transactions are added to the ledger. This protocol is referred to as a consensus mechanism: in a centralized system, the central node could make decisions, whereas in a decentralized system, this protocol

is necessary to ensure that transactions are recorded consistently (Zheng et al., 2017). Various consensus mechanisms exist: Bitcoin and Ethereum currently use *Proof-of-Work*, while Ethereum will be shifting to a *Proof-of-Stake* protocol.

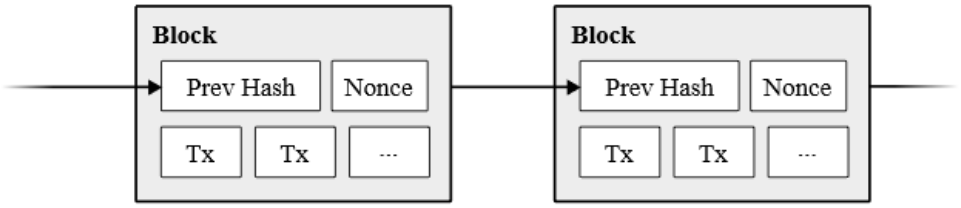

Figure 2: A simple illustration of a blockchain, from Nakamoto (2008).

Because blockchains are decentralized, they avoid the problem that a single node can bring the ledger down - the "single point of failure situation" (Zheng et al., 2017). Because all nodes maintain their own copy of the ledger, for one node to write fraudulent transactions to the ledger, that node would require a majority of the computational power in the network - this is referred to as a 51% attack.

Additionally, blockchains are immutable. Since the ledger is append-only, transactions cannot be edited or tampered with once they have been written to the ledger. This enhances the inherent transparency of such a public ledger.

3.2 Blockchain makes data unique

Data has a few interesting properties that make its economics noteworthy. Firstly, it is costly and time-consuming to produce, yet since it is cheaply replicable, it has infinite supply. Whilst data can be copyrighted, it is difficult to track its destina-

tions and uses once it is released. These properties make the ownership of data an interesting topic and one of vital importance in today's digital economy.

Blockchains allow data to be registered, recorded as unique, and traced to its origins. When each transaction is written to the blockchain, it is 'hashed' and stored on a block of transactions, as illustrated in Figure 2. Hashing is to transform data of arbitrary size into data of fixed size. The hash value is an alphanumeric string that serves as the 'digital fingerprint' of the hashed information. Transactions are written to a block, and this block is again hashed. Nakamoto's (2008) solution involves timestamping each block: this timestamp proves that the data must have existed at that time, and who it was owned by.

Hashing, timestamping, and the transparency of a public ledger make blockchain well-suited to record the ownership of assets, more particularly digital assets - which don't touch the physical world at all and exist only digitally, since all transactions involving the hash of the asset can be easily located and verified.

3.3 Licencing research

Research is a valuable asset - it is a form of intellectual property. UniCoin aims to help academics monetize their research, whilst still maintaining ownership of it. This differs from the traditional model of publishing in an academic journal, where typically copyright of the work is transferred to the journal - although, the growing Open Access movement has resulted in a growing push against this practice (Hoorn and van der Graaf, 2006). How we do this is through licencing the work. A licence is an authorization by the owner of the intellectual property to use their research, whilst not transferring ownership of the IP. This licence can have many special terms: it can be an exclusive licence to the buyer, or it may be valid only for a certain period. Through registering all these details on a blockchain, we treat the

licence as a digital asset, which can be bidded on through the UniCoin platform - this mechanism is investigated further by Maree (2020), who proposes a sealed bid auction combined with an optional Harberger Tax[4] to generate revenue for the academic's university.

The transparency provided by the blockchain provides a key benefit to buyers. Suppose an exclusive licence has been sold. The holder of this licence can see any additional licences, in violation of their agreement with the seller; and prospective buyers can see that the exclusive licence has already been sold, and they are unable to purchase their own licence. Every transaction involving the research that has been registered on the platform can be verified at any point.

How we facilitate the licencing of research is through smart contracts, a key use case of blockchain technology. These are digital contracts which execute predefined actions when specified criteria are met. On the UniCoin platform, they can be used to facilitate the creation of a non-fungible token[5] (NFT) that represents a licence to the use of an academic paper for commercial purposes. Once the prospective buyer of a licence is granted the licence and pays for it, conditions in the smart contract are met, a token is minted, and ownership of that token is granted to the buyer. Ownership of this token needs to be distinguished from ownership of the licence: the token is the digitally native asset which represents the licence, and as such includes all licencing details. The tokenized licence entitles the buyer[6] to a claim on the

[4] "An economic policy that aims to find a balance between pure private ownership and total commons ownership to increase general society welfare and productivity", from Posner and Weyl (2017) as cited by Maree (2020).

[5] A non-fungible token differs from a token like Bitcoin in that tokens are unique and not mutually interchangeable. Further details around the creation of non-fungible tokenized patent licences, as well as the smart contracts to facilitate this, are discussed by Maree (2020).

[6] The buyer can be specified as the only valid licence-holder, as a term of the licence. Alternatively, the holder of the token could be specified to be the valid licence-holder.

underlying licence, which in turn entitles the licence-holder to use the research for commercial applications. This is illustrated below in Figure 3.

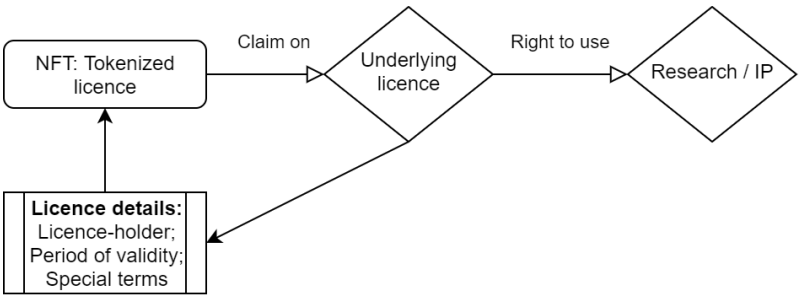

Figure 3: An illustration of the process to tokenize a licence to rights to use research with a commercial application.

Additionally, when an academic publishes their paper, they may wish for it to be open-source and free for research or personal purposes. In the case that a user downloads the paper for research purposes, and has no wish to commercialize it, the licence granted to them could include this detail as well, and the right to use the work for commercial purposes is thus not granted to them. Alternatively, no token at all could be minted, and in the absence of a hash value that contains the details of a licence, the downloaded copy of the paper can be clearly distinguished from a copy that has been provided with the hash value of a legitimate licence. As noted by Zeilinger (2016), these copies are free to be circulated, as they attest to the value of the original work, and can potentially impact the value of a licence.

UniCoin uses the strengths of blockchain technology to create a marketplace for research. We take this further, and trace research outputs to their origins in terms of the research that influenced it, in a manner that fairly recognizes all parties in relation to their contribution, to ensure licencing revenues are distributed fairly.

4 Attributing credit between cited papers

4.1 Methodology

Many approaches to attributing credit between coauthors can be taken. The most effective, but least practical to implement in a credit diffusion mechanism, would be for authors to actually indicate their respective contributions - no misallocation of credit between primary and secondary contributors would result.

The second major source of knowledge that deserves to have credit attributed is the body of work that is cited within a paper. Just as authors contribute different proportions to the publication of a paper, so too do the papers cited within that paper - a paper could reference the framework supporting a new theory, or simply be part of the motivation to conducting the research in question (Fang, 2018).

Fang (2018) represents this in the following form: M authors contribute some proportion α_i, and N references each contribute some proportion β_j, with

$$\sum_{i=1}^{M} \alpha_i + \sum_{j=1}^{N} \beta_j = 1$$

The framework on how to split credit between authors and references has not been tested empirically. The approach suggested by Fang (2018) is to use content citation analysis to determine how a reference contributes to a paper - this can influence how much credit that reference receives. Content citation analysis uses the sentence surrounding a citation to classify it in certain predefined categories - it can identify not only how a citation is used, but also *why* (Ding et al., 2014). All references can be categorized like this, and then using expert judgement and machine learning techniques, the distribution of references across various categories can be used to determine what contribution the authors have made over the references, and hence determine the size of $\sum \alpha_i$. To apply the proposed solution is not a trivial task, and

for this reason will not be explored further in this thesis.

In building a mechanism to distribute credit across a citation network, again one *ad hoc* solution would be for authors to simply specify what proportion should be allocated to references. Authors could also specify a third group of individuals, who are not listed as authors or references, but did also contribute to the paper - such as lab assistants or colleagues who commented on the work - and the above proportions could be rescaled, where γ_k represents the proportion allocated to the k-th non-listed contributor, so that

$$\sum_{i=1}^{M} \alpha_i + \sum_{j=1}^{N} \beta_j + \sum_{k=1}^{O} \gamma_k = 1$$

This third group of contributors is not one we can easily integrate into our framework - it would require either analysis of informal collaboration, as in Rose and Georg (2018), or a manual selection of non-listed contributors by the authors. So, we don't delve into this this group further in this thesis, and focus on finding a better approach to allocate credit between the first two groups, using the structure of the network. So, we use centrality scores to influence the division between authors and cited papers.

This division will be influenced by the network of papers surrounding a paper, so the split is determined for each paper of interest. Calling this paper *Paper 0*, we define a paper's generation in relation to its geodesic distance[7] to Paper 0. So, *Generation 1* consists of all of the papers which cite Paper 0. *Generation 2* consists of all papers which cite the papers in *Generation 1*, and so on. *Generation -1* consists of all papers which have been cited by Paper 0. Forward and backward generations are discussed in more detail by Hu et al. (2011).

A paper which may appear in multiple generations is classified in the closest gen-

[7] The number of edges in the shortest path between two nodes.

eration to Paper 0 - so, a paper cited by Paper 0 as well as by another paper in *Generation -1* falls in *Generation -1*, not *Generation -2*.

Considering the example citation network in Figure 1 on Page 6, if we take $P4$ to be the paper of interest, then *Generation 1* consists of $P1$ and $P3$, and *Generation -1* consists of $P5$.

We can compute the centrality scores for each paper in the network, and compare the scores in *Generation -1* with the score of Paper 0. A higher score indicates a more influential paper - it has received more citations, or more citations from other influential papers. This higher score indicates that in this network, this paper is more influential, and should receive a larger share of credit.

The centrality scores for *Generation -1* and Paper 0 are discounted. In *Generation -1*, papers from a wide range of publication years can appear. Papers which are a lot older have an unfair advantage, as they are more likely to have attained many more citations than younger papers. To counteract this, the centrality scores are multiplied by $\delta^{(2019-T_i)}$ where δ is some discount factor $(0 < \delta < 1)$[8] and T_i is the year of publication of paper i. We denote these discounted centrality scores by Y_i for $i = 0, 1, 2, ..., N$ where papers $1, 2, ..., N$ fall in *Generation -1*.

In this paper, we compare the following centrality measures: eigenvector centrality, Katz centrality, PageRank, and in-degree centrality, which simply counts the number of citations a paper has received, and normalizes this.

To decide how much credit is retained with Paper 0, and how much credit is passed to *Generation -1*, the average discounted centrality score is computed for *Generation -1*. Now, we have two time-adjusted centrality scores, that are in the same scale (of one paper), and can be compared.

[8] We use a discount factor of 0.98. Varying this factor between 0.97 and 0.99 changes the proportion allocated to Paper 0 by less than 1.5%.

Calling this average discounted centrality score $\overline{Y_{-1}}$, to Paper 0 we attribute

$$\sum_{i=1}^{M} \alpha_i = \frac{Y_0}{Y_0 + \overline{Y_{-1}}}$$

The remainder is allocated to *Generation -1*. The centrality scores of each paper influence the proportion they receive. Of the credit that is passed to this generation, paper i will receive a share

$$\theta_i = \frac{Y_i}{\sum_{i=1}^{N} Y_i}$$

Hence, the total amount of credit that paper i in *Generation -1* will receive is

$$\beta_i = \theta_i \times \frac{\overline{Y_{-1}}}{Y_0 + \overline{Y_{-1}}} = \frac{Y_i}{\sum_{i=1}^{N} Y_i} \times \frac{\overline{Y_{-1}}}{Y_0 + \overline{Y_{-1}}} = \frac{\frac{1}{n} Y_i}{Y_0 + \overline{Y_{-1}}}$$

4.2 Data

Data were collected from Scopus, Elsevier's abstract and citation database, using the `pybliometrics` Python package from Rose and Kitchin (2019).

There are three potential approaches for defining the scope and breadth of a citation network, that is to be built from data collected from Scopus. The first is to consider the entire data universe - that is, all 69 million records on Scopus' database, and map edges between these nodes and compute the resulting centralities for each node. This approach is obviously the most rigorous and will result in the most true representation of the network. However, this approach is not considered due to computational limitations.

The second approach, a top-down approach, is to define the network based on a collection of journals and a collection of years of publication. Data from a top percentile of journals could be collected for all papers published over a defined period - e.g. the top 20 economics journals and all papers published in the last 30 years. This

is more computationally feasible than the prior approach, and ultimately the depth and breadth of the network can be carefully defined to control for this. Another advantage of this approach is that by defining one large network, when a smaller sub-network is viewed, the centralities of the nodes viewed will be in relation to the entire network, and these centralities will not change depending on what subsection of the network is being examined. One disadvantage, however, is that through controlling which journals' data is included in the network, influential papers or books which are not published in those journals will be excluded from the network completely, ignoring that they may in fact be a critical part of the network.

A third approach, which overcomes this disadvantage, is to construct the network in a bottom-up fashion, whereby a network is constructed directly in relation to a paper of interest. A paper of interest is selected, and from here all data relating to papers in a predefined number of generation is collected. Rather than limiting the amount of data by journals and years, the amount of data is limited in relation to distance from the paper of interest. So, *Generation 1, -1* and *-2* can be collected, as well as all papers collected directly to these generations. To make a more complete and 'true' depiction of the network, data from more generations can be collected. This is the approach we use for data collection in this thesis.

4.3 Results

Inspired by Tol (2011), we choose to illustrate the results through looking at three papers by Andrei Shleifer, Harvard University Professor of Economics. To observe how the results differ by number of authors, number of citations, and age of the paper, we use the following three papers[9]:

[9]Notice how the authors are listed alphabetically - so, if rank weights were used to account for coauthorship, Shleifer would receive the lowest share in each case.

- Example 1: A widely cited 2008 paper from the Journal of Economic Literature, *'The economic consequences of legal origins'*, which has received 1224 citations in Scopus at the time of writing, and is authored by La Porta, Lopez-de-Silanes, and Shleifer.
- Example 2: A less widely cited 2014 paper from the Journal of Economic Perspectives, *'Informality and development'*, which has received 171 citations in Scopus at the time of writing, and is authored by La Porta and Shleifer.
- Example 3: Lastly, a young paper which has had not much time to receive many citations, *'Extrapolation and Bubbles'*, from the Journal of Financial Economics, published in 2018, which has received thus far 17 citations in Scopus and is authored by Barberis, Greenwood, Jin, and Shleifer.

We construct the networks and calculate the centrality scores for all papers in each network. The networks differ significantly in size: Example 1 consists of $146,317$ nodes, Example 2 of $22,449$ nodes, and Example 3 of $13,054$ nodes.

Some network illustrations are shown: these are not the full networks, which are too large to visualize, but consist of all nodes in Generations 0, -1, and -2. Each network is shown below: nodes are scaled by in-degree (number of citations in the subnetwork), and coloured by generation. *Generation 0* is green, *Generation -1* is orange, and *Generation -2* is lilac.

Network visualization is done with Gephi from Bastian et al. (2009), and nodes are positioned using the Force Atlas force-directed layout.

Since these are still quite large visualizations, more interactive versions are online at https://lukemeiklejohn.github.io/#thesis

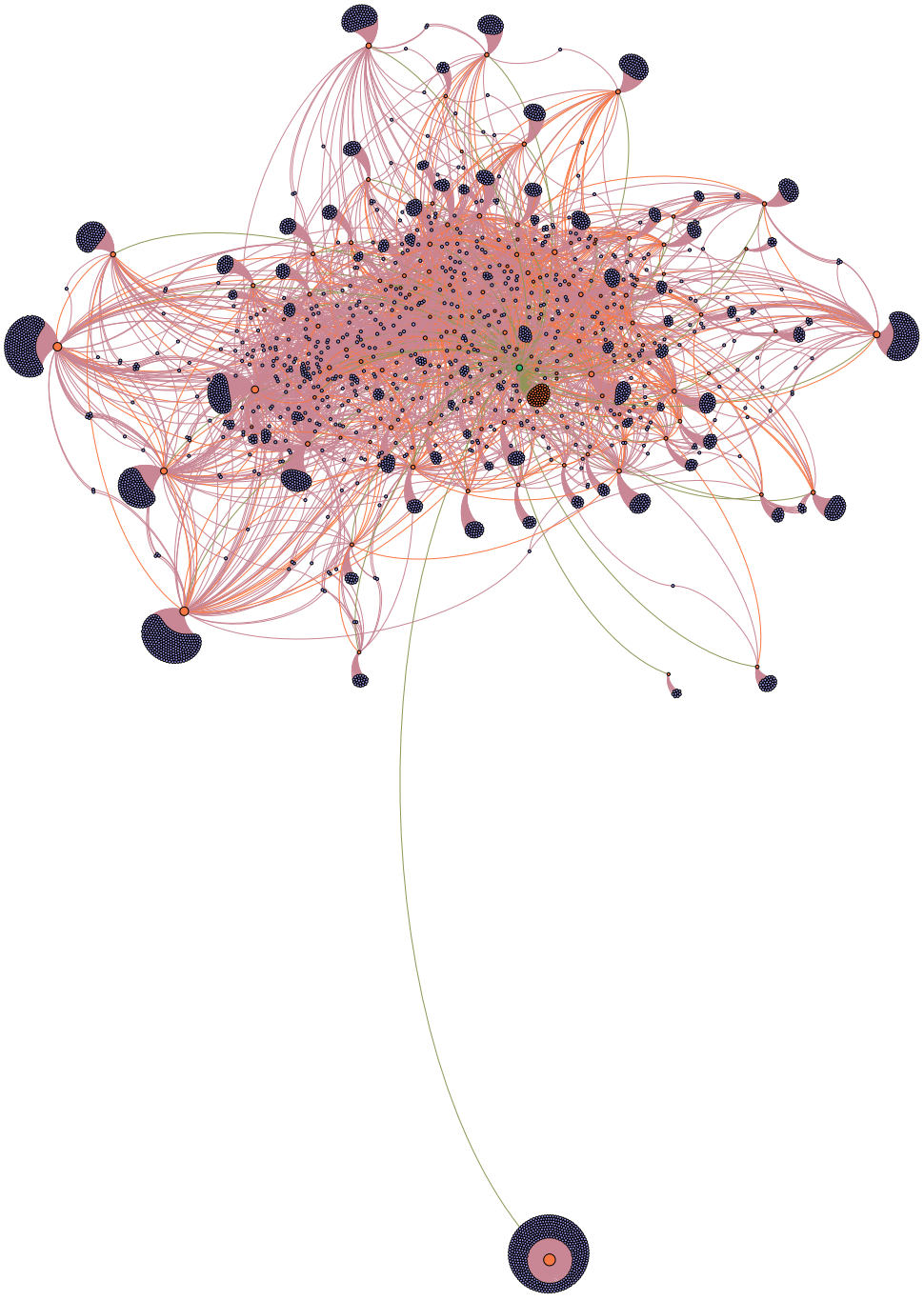

Figure 4: Generations 0, -1, and -2 for Example 1.

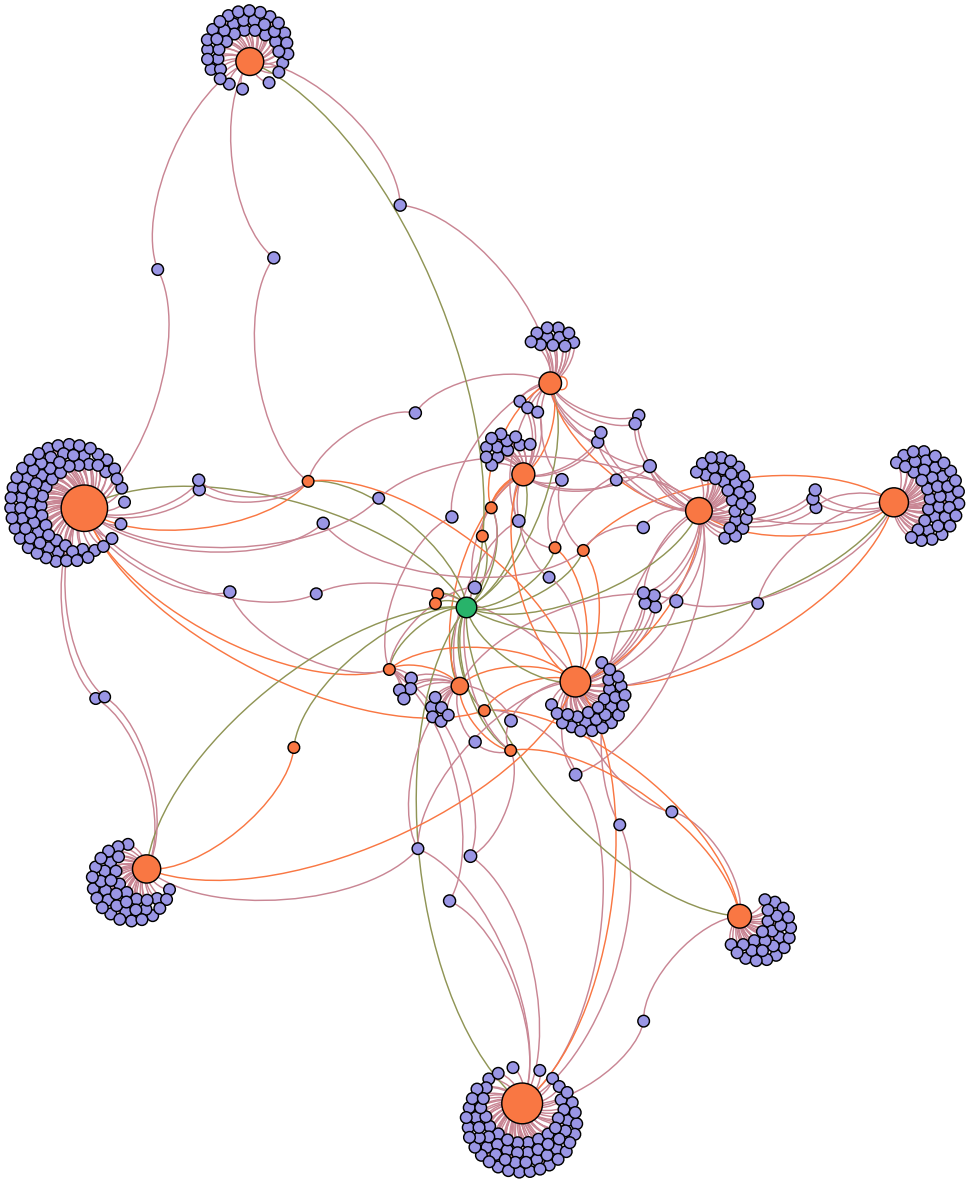

Figure 5: Generations 0, -1, and -2 for Example 2.

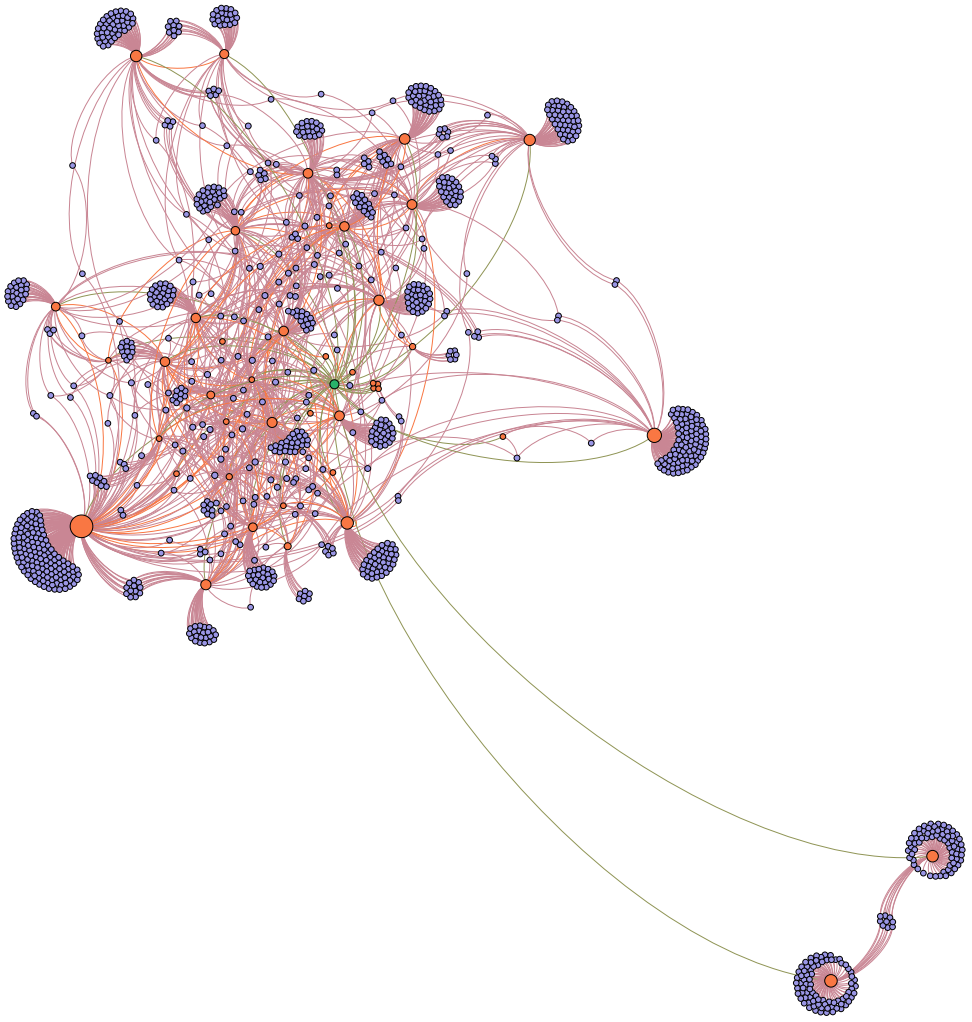

Figure 6: Generations 0, -1, and -2 for Example 3.

In the cases of Figure 4 and Figure 6, there are cases of small communities which are noticeably distinct from the rest of the network. In the case of Example 3, there are two papers in *Generation -1* which have resulted in three clusters of papers separate from the network.

These two papers are:

- *'Simultaneous modeling of visual saliency and value computation improves predictions of economic choice'* by Towal, Mormann, and Kock
- *'Neuroeconomic foundations of economic choice-recent advances'* by Fehr and Rangel

Both papers are significantly more scientific than the rest of the papers cited by Paper 0, which are more traditionally economic - and so, it makes sense for these communities of papers to be separate from the rest of the network.

The centrality scores for each of the Paper 0s are as follows:

	Katz centrality	In-degree centrality	Eigenvector centrality	PageRank
Example 1	0.06590	0.00119	0.00441	0.00016
Example 2	0.03163	0.00098	0.00019	0.00011
Example 3	0.13642	0.00329	0.11004	0.00095

Table 1: Centrality scores for the paper of interest in each of the named examples.

The varying scales are interesting to note (in Example 3, the paper of interest has the highest centrality scores in the table above, due to it being in a much smaller network); however, what is more valuable is to compare these scores in relation to the rest of each network.

As an example, we can view for Example 3 an excerpt of the rankings in terms of their discounted centrality scores below. These rankings are constructed based on the most influential papers which are cited by Paper 0. More information on each of the papers in Table 2 are given in Appendix A.

Paper ID	PageRank	In-degree	Katz	eigenvector
7	1	1	1	3
10	2	2	4	•
37	3	4	7	•
14	4	6	•	•
16	5	3	2	2
13	6	5	•	•
18	7	7	•	•
41	•	•	3	4
17	•	•	5	1
3	•	•	6	5
33	•	•	•	6
22	•	•	•	7

Table 2: An excerpt of the ordinal rankings of papers in terms of centrality scores.

Initial results reveal very similar rankings for the PageRank and in-degree centrality measures. Katz centrality appears somewhat similar, whilst eigenvector centrality has the least similar rankings to the other three measures.

Whilst the rankings in terms of centrality are interesting to compare, it is more telling to compare the allocations of credit across the network. The allocations given to the Paper 0s are given in Table 3 below. Note, that this is the amount of credit allocated to the *paper*, not the author. This allocation will be split further to account for coauthorship of each paper. We allocate credit at a paper level and then correct for multiple authorship, rather than allocating credit at a researcher level, which causes mis-allocation of credit (Wang et al., 2016).

	PageRank	In-degree centrality	Katz	Eigenvector centrality
Example 1	71.02%	85.84%	95.53%	99.37%
Example 2	23.75%	53.95%	74.28%	95.74%
Example 3	43.79%	49.21%	84.55%	97.75%

Table 3: Credit allocations given to the central paper in each of the examples, guided by the centrality score named.

Two immediate points of interest are that for Example 1, with 1224 citations, all centrality measures guide a majority of the credit to the paper of interest. This is a desirable trait! Secondly, that eigenvector centrality appears to always result in almost all credit being attributed to the paper of interest: this bias is a weakness, and in directed acyclic graphs such as citation networks, eigenvector centrality results in the problem of zero centrality. Nodes with no in-edges have a centrality of zero, and any node that has only one in-edge from that node also has a centrality of zero. This problem is made clear in Table 3, where many nodes have very small centrality, and hence eigenvector centrality is not suitable for use in a platform such as UniCoin. Katz Centrality is a useful adaptation that can be used instead.

Where the ordinal rankings by PageRank and in-degree centrality are almost identical, their credit allocations vary significantly: only Example 3 has somewhat similar credit allocations for the paper of interest. Katz centrality still results in larger credit allocations for each of the papers.

For each of the networks, we plot below in Figures 7, 8, and 9 the credit allocations for each of the papers in the respective *Generation -1*. The full data are available here, as mentioned in Appendix B.

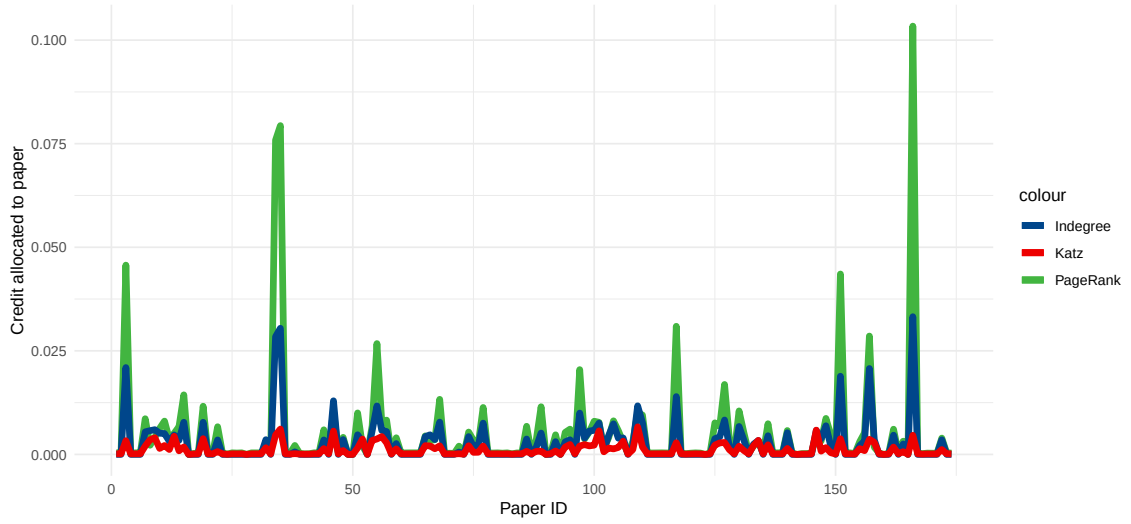

Figure 7: Credit allocations for each of the papers in Example 1's *Generation -1*.

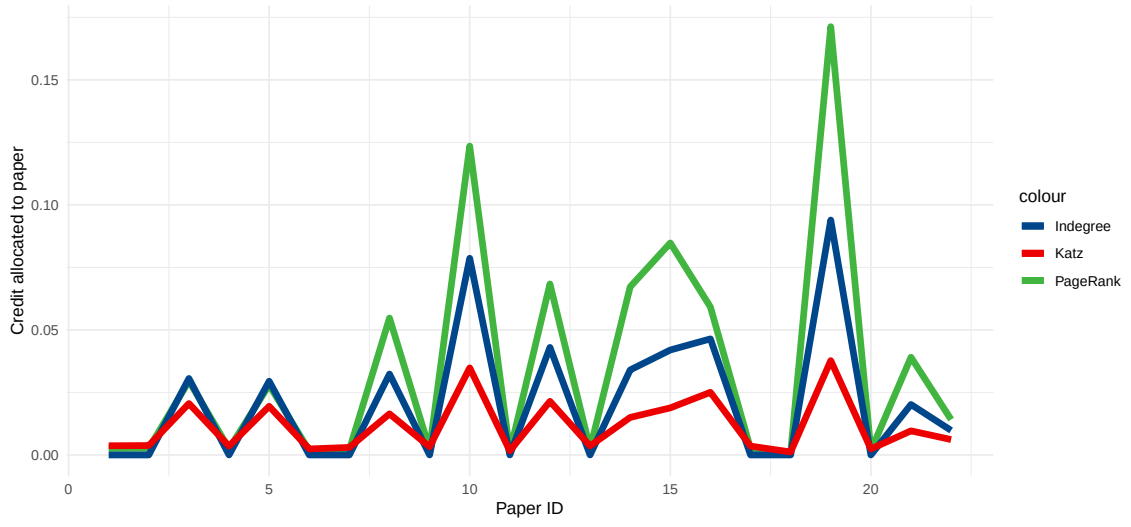

Figure 8: Credit allocations for each of the papers in Example 2's *Generation -1*.

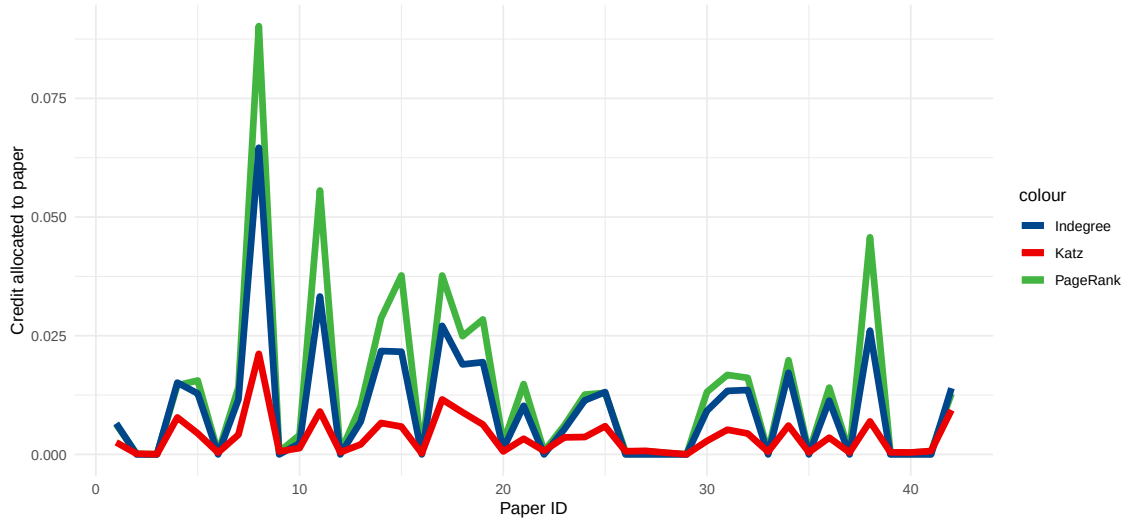

Figure 9: Credit allocations for each of the papers in Example 3's *Generation -1*.

These three graphs essentially show the other side of Table 3: while PageRank results in the lowest share of credit for the paper of interest out of the three measures, it is the most generous method for attributing credit to cited papers. Likewise, the opposite is true for Katz centrality, and in-degree centrality consistently allocates credit somewhere between the other two measures. Out of these three measures, it could be said that in-degree centrality is thus the most 'fair' - neither too generous to cited papers, nor too generous to the paper of interest. By the principle of Occam's razor, simply using number of citations as the mechanism to diffuse credit may thus in fact be the best way.

If number of citations is a good proxy for measuring impact in the network, the PageRank results are slightly inconsistent. For Example 2, the paper of interest has received 171 citations, while the average paper in *Generation -1* has received 21 citations. The paper of interest only receives 23.75% of the credit here, whilst in Example 3, the paper of interest has only received 17 citations compared to the

average paper in *Generation -1* with 35 citations, and receives 43.79% of the credit. This dilemma is repeated with the results from using Katz centrality.

One limitation of these results is the size of each of the generations of cited papers. Whilst the papers of interest in Example 1 and 3 cite 174 and 42 papers respectively, the paper of interest in Example 2 only cites 22 papers. Thus, this paper is only connected directly to 39 papers, and this may explain the anomalies with PageRank and Katz centrality.

In empirical work by Litvak et al. (2007), a relationship between the tails of a website's PageRank and its in-degree was found - that the distributions of these centrality scores follow power laws with the same exponent. A power law says that $\Pr[X > x] \approx x^{-\alpha}$. This result is in line with our results, which show that for cited papers, the *manner* in which credit is allocated is very similar, only the scale differs. PageRank not only grows with in-degree, but there is a significant correlation between the two measures - Fortunato et al. (2006) devises a method to approximate PageRank from in-degree for the purposes of ranking websites, as in the results of an internet search query. This approximation for the purposes of ranking bears a striking resemblance to our results - recall that our ranking of papers in Table 2 for the results from in-degree was incredibly similar to the ranking from PageRank.

Comparing Katz centrality with some other centrality measures through similar empirical work, Bloch et al. (2016) found that with a low decay factor, Katz centrality acts as a very similar measure of centrality to degree. For very small parameter values, Katz centrality scores are mostly influenced by short paths, such as paths of length one - that is, in-degree. As the decay factor increases, Katz centrality incorporates more information from longer paths rather than just a node's immediate neighbours, and starts to differ more significantly. We did not experiment with variations of this parameter, but hypothesise that if we were to decrease it, the

allocations would be more similar to the allocations by in-degree centrality, and less similar if we were to increase it. Tuning the Katz centrality decay parameter for our application is one potential area of future research.

5 Conclusion

This thesis uses citation networks to develop a mechanism to fairly attribute credit amongst sources of knowledge that result in the creation of an academic publication. We share credit between the authors of a paper, and all papers cited, as well as allow for a third group of contributors: those who traditionally would not receive credit, such as lab assistants or informal collaborators. Through three empirical case studies based on papers by Andrei Shleifer, we find that Katz centrality and PageRank are useful network centrality measures that can be used for this credit diffusion mechanism, and that surprisingly, in-degree centrality is also a very strong contender for this application.

All three measures result in slightly different allocations of credit across the network, and whilst each of the results presented can be considered fair, it is a more esoteric task to ask which of these allocations is *most* fair: this is a task that would need to be conducted with a more finely tuned knowledge of the subject domain and a knowledge of how the paper of interest contributes to the domain in comparison to how the cited papers do. Indeed, this brings us back to the approach suggested by Fang (2018) using content citation analysis, machine learning, and expert judgement, as mentioned in Section 4.1.

We discussed why blockchain technology should be harnessed for an academic licencing application: it allows academics to retain ownership of their intellectual property, while using smart contracts to issue NFTs that function as licences. This licencing platform can be combined with the credit diffusion mechanism discussed to ensure that licencing revenue is distributed fairly in a manner that recognizes all contributors in proportion to their contribution.

Future work can involve integrating this credit diffusion mechanism into UniCoin: for this thesis, we used data that were already uploaded onto Scopus, and because of

this, mapping edges from a paper to its cited papers was a fairly easy task. To add value, the platform should be able to scrape any arbitrary paper and extract the cited papers from it, before linking those papers to their Scopus data and constructing the citation network. The visualisation element can also be integrated into the UniCoin website: what we have shown in this thesis serves as a proof of concept, but in the final product it should be possible for any paper to be selected or uploaded, and the citation network constructed as well as visualized.

A Full paper details from Table 2

Note: Not all of this data was available/correct on Scopus. Some manual data correction was done below, but this was not required for the thesis and was done purely to provide complete information here.

Paper ID	EID[10]	Paper Title	Authors	Publication	Year
7	2-s2.0-4043089417	Hedge funds and the technology bubble	Brunnermeier & Nagel	Journal of Finance	2004
10	2-s2.0-84877974019	What have they been thinking? Homebuyer behavior in hot and cold markets	Case, Shiller, & Thompson	Brookings Papers on Economic Activity	2012
37	2-s2.0-0001575872	Asset bubbles and overlapping generations	Tirole	Econometrica	1985
14	2-s2.0-84904437662	Two pillars of asset pricing	Fama	Am. Econ. Rev.	2014
16	2-s2.0-0004320711	The Great Crash: 1929	Galbraith	None	1954
13	2-s2.0-84977712440	Positive Feedback Investment Strategies and Destabilizing Rational Speculation	De Long, Shleifer, Summers, & Waldmann	The Journal of Finance	1990
18	2-s2.0-85021799760	An extrapolative model of house price dynamics	Glaeser & Nathanson	Journal of Financial Economics	2017
41	2-s2.0-84894322066	Bubbles, crises, and heterogeneous beliefs	Xiong	Handbook for Systemic Risk	2013
17	2-s2.0-84969217968	No-Bubble Condition: Model-Free Tests in Housing Markets	Giglio, Maggiori, & Stroebel	Econometrica	2016
3	2-s2.0-0004058553	Lombard Street: A Description of the Money Market	Bagehot	None	1873
33	2-s2.0-0004179594	Irrational Exuberance	Shiller	None	2000
22	2-s2.0-85049295829	Hoard Behavior and Commodity Bubbles	Hong, de Paula, & Singh	NBER Working Paper No. 20974.	2015

[10]These are the unique identifiers assigned to academic works in the Scopus database.

B Data and code

All data and code are available on the following GitHub repository:

https://github.com/lukemeiklejohn/HowToAttributeCredit

There are three folders, namely:

- Figures, containing all named figures from this document.
- Code, containing Python code (in the form of Jupyter notebooks) for Scopus scraping and data processing.
- Data, containing processed data from Excel, as well as the Gephi network files.

References

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. Wanted: The key player. Econometrica, 74(5):1403–1417.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

Bloch, F., Jackson, M. O., and Tebaldi, P. (2016). Centrality measures in networks.

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., and Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. Journal of the Association for Information Science and Technology, 65(9):1820–1833.

Drescher, D. (2017). Seeing the big picture. In Blockchain Basics, pages 9–17. Apress.

Egghe, L. (2006). Theory and practise of the g-index. Scientometrics, 69(1):131–152.

Fang, H. (2018). A discussion of citations from the perspective of the contribution of the cited paper to the citing paper. Journal of the Association for Information Science and Technology, 69(12):1513–1520.

Fortunato, S., Boguñá, M., Flammini, A., and Menczer, F. (2006). Approximating PageRank from in-degree. In Algorithms and Models for the Web-Graph, pages 59–71. Springer Berlin Heidelberg.

Gilbert, G. N. (1977). Referencing as persuasion. Social Studies of Science, 7(1):113–122.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, 102(46):16569–16572.

Hoorn, E. and van der Graaf, M. (2006). Copyright issues in open access research journals. D-Lib Magazine, 12(2).

Hu, X., Rousseau, R., and Chen, J. (2011). On the definition of forward and backward citation generations. Journal of Informetrics, 5(1):27–36.

Jackson, M. (2010). Social and economic networks. Princeton University Press.

Kim, J. and Diesner, J. (2014). A network-based approach to coauthorship credit allocation. Scientometrics, 101(1):587–602.

Kostoff, R. N. (1998). The use and misuse of citation analysis in research evaluation. Scientometrics, 43(1):27–43.

Litvak, N., Scheinhardt, W. R. W., and Volkovich, Y. (2007). In-degree and pagerank: Why do they follow similar power laws? Internet Math., 4(2-3):175–198.

Maree, C. (2020). A marketplace for ideas: improving allocation efficiency in commercially viable academic publications. Master's thesis, University of Cape Town.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

Newman, M. (2010). Networks: An Introduction. Oxford University Press.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Perry, M. and Reny, P. J. (2016). How to count citations if you must. American Economic Review, 106(9):2722–2741.

Portenoy, J., Hullman, J., and West, J. D. (2017). Leveraging citation networks to visualize scholarly influence over time. Frontiers in Research Metrics and Analytics, 2.

Posner, E. A. and Weyl, E. G. (2017). Property is only another name for monopoly. Journal of Legal Analysis, 9(1):51–123.

Radicchi, F., Fortunato, S., Markines, B., and Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. Physical Review E, 80(5).

Rose, M. E. and Georg, C.-P. (2018). What 5,000 acknowledgements tell us about informal collaboration in financial economics. SSRN Electronic Journal.

Rose, M. E. and Kitchin, J. R. (2019). pybliometrics: Scriptable bibliometrics using a python interface to scopus. SoftwareX, 10:100263.

Rusinowska, A., Berghammer, R., Swart, H. D., and Grabisch, M. (2011). Social networks: Prestige, centrality, and influence. In Relational and Algebraic Methods in Computer Science, pages 22–39. Springer Berlin Heidelberg.

Stallings, J., Vance, E., Yang, J., Vannier, M. W., Liang, J., Pang, L., Dai, L., Ye, I., and Wang, G. (2013). Determining scientific impact using a collaboration index. Proceedings of the National Academy of Sciences, 110(24):9680–9685.

Tol, R. S. J. (2011). Credit where credit's due: accounting for co-authorship in citation counts. Scientometrics, 89(1).

Wang, H., Shen, H.-W., and Cheng, X.-Q. (2016). Scientific credit diffusion: Researcher level or paper level? Scientometrics, 109(2):827–837.

West, J. D., Bergstrom, T. C., and Bergstrom, C. T. (2010). The Eigenfactor Metrics™: A network approach to assessing scholarly journals. College & Research Libraries, 71(3):236–244.

Zeilinger, M. (2016). Digital art as 'monetised graphics': Enforcing intellectual property on the blockchain. Philosophy & Technology, 31(1):15–41.

Zheng, Z., Xie, S., Dai, H., Chen, X., and Wang, H. (2017). An overview of blockchain technology: Architecture, consensus, and future trends. In 2017 IEEE International Congress on Big Data (BigData Congress). IEEE.