



Universidad de Alcalá

Programa de doctorado en ciencias

Doctorate programme of sciences

Applications of graph theory to wireless networks and opinion analysis

Thesis submitted by Marino Tejedor Romero

Supervisor **David Orden Martín**
Departamento de Física y Matemáticas
Department of Physics and Mathematics

Supervisor **Iván Marsá Maestre**
Departamento de Automática
Department of Automatics

March, 2023

Abstract

Graph theory is an important field of discrete mathematics. It has been widely used lately thanks to the rise of data science as a discipline, since it is useful for structuring, analyzing and generating data through models. The objective of this thesis is to employ graph theory for the optimization of wireless networks and opinion analysis.

The first set of contributions is related to graph theory applied to wireless networks. Performance of these networks depends on the appropriate distribution of their frequency channels in a shared physical space. In order to optimize these networks, several techniques had been proposed, from heuristics like simulated annealing to automatic negotiation. All these techniques require a model behind it. We developed a Wi-Fi network model using geometric graphs. Nodes represent devices, both clients and access points, and edges represent desired or undesired signal between two devices. These graphs are geometric. Consequently, vertices have a position in the space, and edges have a length. Including these parameters and a propagation model, we are able to simulate wireless networks and contribute to optimize them. Using this graph-based model, we studied the effect of cochannel-interference in Wi-Fi 4 networks and showed that channel-bonding offers a performance gain when used in regions where 13 channels or more are available.

On the other hand, we have applied graph theory to opinion analysis along the line of *SensoGraph*, which is a method enabling efficient sensory analysis of a set of items by means of proximity graphs, allowing to handle large input datasets. In addition, we have developed a new opinion analysis method, which uses manual edge assignment and distances in a graph to study pairwise similarity between different samples.

Finally, we have explored themes outside of graph theory but still inside the frame of mathematics applied to a telematics engineering problem. A novel e-voting system based on mixnets, Shamir's secret sharing and finite fields is proposed. This proposal offers a new verification system and complies with the essential properties of voting systems.

Resumen

La teoría de grafos es una rama importante dentro de la matemática discreta. Su uso ha aumentado recientemente dada la conveniencia de los grafos para estructurar datos, para analizarlos y para generarlos a través de modelos. El objetivo de esta tesis es aplicar teoría de grafos a la optimización de redes inalámbricas y al análisis de opinión.

El primer conjunto de contribuciones de esta tesis versa sobre la aplicación de teoría de grafos a redes inalámbricas. El rendimiento de estas redes depende de la correcta distribución de canales de frecuencia en un espacio compartido. Para optimizar estas redes se proponen diferentes técnicas, desde la aplicación de heurísticas como *simulated annealing* a la negociación automática. Cualquiera de estas técnicas requiere un modelo teórico de la red inalámbrica en cuestión. Nuestro modelo de redes Wi-Fi utiliza grafos geométricos para este propósito. Los vértices representan los dispositivos de la red, sean clientes o puntos de acceso, mientras que las aristas representan las señales entre dichos dispositivos. Estos grafos son de tipo geométrico, por lo que los vértices tienen posición en el espacio, y las aristas tienen longitud. Con esta estructura y la aplicación de un modelo de propagación y de uso, podemos simular redes inalámbricas y contribuir a su optimización. Usando dicho modelo basado en grafos, hemos estudiado el efecto de la interferencia cocanal en redes Wi-Fi 4 y mostramos una mejora de rendimiento asociada a la técnica de *channel bonding* cuando se usa en regiones donde hay por lo menos 13 canales disponibles.

Por otra parte, en esta tesis doctoral hemos aplicado teoría de grafos al análisis de opinión dentro de la línea de investigación de *SensoGraph*, un método con el que se realiza un análisis de opinión sobre un conjunto de elementos usando grafos de proximidad, lo que permite manejar grandes conjuntos de datos. Además, hemos desarrollado un método de análisis de opinión que emplea la asignación manual de aristas y distancias en un grafo para estudiar la similaridad entre las muestras dos a dos.

Adicionalmente, se han explorado otros temas sin relación con los grafos, pero que entran dentro de la aplicación de las matemáticas a un problema de la ingeniería telemática. Se ha desarrollado un sistema de votación electrónica basado en mixnets, secreto compartido de Shamir y cuerpos finitos. Dicha propuesta ofrece un sistema de verificación numérico novedoso a la vez que mantiene las propiedades esenciales de los sistemas de votación.

Acknowledgements

This thesis has only been possible thanks to the continuous support of many people I have found in my life, and a lucky pile of circumstances. Hard work is only one of the necessary ingredients. This is why I have no physical space to say thanks to all the people and resources that made this possible.

First of all, I want to thank my **advisors in Universidad de Alcalá** that guided me and helped me inside academia: **Iván Marsá Maestre, David Orden Martín and José Manuel Giménez Guzmán**. Without their counseling, I would be completely lost in all fields of my PhD and previous activities in the academia: writing, teaching, bureaucracy, etc. Thanks a lot for their patience, their time and their efforts, and for inviting me to collaborate in their projects when I was still studying my degree in Telematics Engineering.

I would also like to thanks my PhD colleagues **David Herranz Oliveros** and **Alejandra Martinez Moraian** for the moments we shared.

Special thanks to **Javier Junquera Sánchez** and **José Raúl Durán Díaz**, among other teachers in the **Master's Degree in Cybersecurity**, who inspired me and supported me in my e-voting research.

A very important part of my doctorate has been the **visit to the Delft University of Technology** for three months in 2022. I want to thank **Pradeep K. Murukanniah** and **Catholijn M. Jonker** for making it possible and for their guidance when collaborating in TU Delft and when I was living in the Netherlands.

It is also important to mention other **coauthors: Encarnación Fernández-Fernández, Luis Cruz-Piris, Susel Fernandez, Jacob Lahne and Katherine Phetxumphou**. It has been a pleasure to research with them and exchange ideas.

I want to thank the whole institution of **Universidad de Alcalá** for all the resources they have invested in my formation: degree, master's degree and PhD. I am grateful and fortunate to live in a welfare state, where I had access to public education with public funding.

At the same time, there are countless resources I have been able to use for free: Several distributions of **Linux**, **LaTeX** (and the current template used, **CleanThesis**),

Visual Studio Code, LibreOffice, Firefox, Thunderbird, Python (and the libraries I used, like **NumPy, SciPy, NetworkX, Numba, Matplotlib, Plotly, Flask**, etc.), and many other unknown heroic packages that make our work possible. All these resources I used are FOSS: Free Open Source Software. Thanks to the **open source community** for **allowing everyone to make science** and contribute back. Honorable mention for **Javier Ruiz**, a friend who, like me, enjoys math videos, and introduced me to Manim, a Python library created by 3Blue1Brown (Grant Sanderson) and then forked by the community. Manim has helped me greatly when creating animations for both teaching and presenting my research findings.

Finally, I want to thank my close ones, family and friends, for all the support and the good moments we have shared so far. Especially my **mom**, a tireless supporter every single moment, and my closest friends **Aarón, Luis Miguel, Laura, Iván and Fernando**. Thank you all.

In conclusion, as the saying goes, we are **standing on the shoulders of titans**, and I feel grateful and fortunate for all the previous knowledge, all the resources, and every person that helped me in my way. Let's keep it that way; let's promote a communitary effort.

Contents

1	Introduction	1
1.1	Graphs applied to wireless networks	3
1.2	Graphs applied to opinion analysis	5
1.3	Electronic voting	7
2	IEEE 802.11 graph models	11
3	Channel Bonding in Wi-Fi 4	29
4	Geometric projective mapping for opinion analysis	41
5	The free-linking task	53
6	Distributed Remote Secure E-Voting	65
7	Conclusions	85
7.1	Conclusiones	86
7.2	Future Work	88
7.3	Trabajo futuro	89
	Bibliography	91

Introduction

” *You said science was about admitting what we don’t know.*

— **Murphy Cooper**
(Interstellar)

Graph theory is an important field of discrete mathematics. It has been a particularly powerful tool paired with many different fields of computer science. For instance, graphs are a regularly used data structure in many programming languages. It is not only used inside computer science, but also as part of data science. Graph theory tools allow data scientists to analyze data with ease and precision [Cos21]. Graphs are also used to understand the world through simplified models using vertices and edges.

Graphs are mathematical structures composed by vertices (also called nodes), and edges (also called links). Formally, a graph is an ordered pair $G = (V, E)$, where V is the set of vertices, and E is the set of edges between elements in V , that is, $E \subseteq \{\{i, j\} | i, j \in V \text{ and } i \neq j\}$. To be more precise, this is the definition of an undirected, abstract, unweighted and simple graph. There are more complex kinds of graphs which include more information, such as:

1. **Directed graphs:** Graphs whose edges are directed from one vertex to another. This means that edges are not just pairs $\{i, j\}$, but ordered pairs (i, j) . As a consequence, given two vertices i and j , there could be: no edge between them, the edge (i, j) , the edge (j, i) , or both at the same time.
2. **Geometric graphs:** Graphs whose vertices have a position, typically two- or three-dimensional. As a result, edges have a length determined by the position of its two vertices.
3. **Weighted graphs:** Graphs whose edges are assigned a weight. In this type of graph, edges are not just existing or not, since existing edges may have different weights.

4. **Multigraphs:** Graphs where E is a multiset of edges. This allows the existence of parallel edges, connecting the same two vertices.

These types are often mixed, depending on the particular structure of the problem we are considering, and also enhanced with additional ad-hoc attributes associated to its vertices or edges. This is why graphs are a flexible framework for many different problems. Most of the graphs covered in following sections will be geometric weighted graphs.

Graph theory comes not only with this concept, but with a powerful toolkit. It includes all kinds of metrics, algorithms and processes. To mention some important examples: clustering, outlier detection, graph coloring, graph layouts, graph visualization, distance-related metrics, centrality, etc. One notable example is the concept of proximity graphs. These graphs represent relations determined by a given neighbourhood condition. Depending on that condition, the existence or not of an edge is determined. The Gabriel Graph [GS69] is the proximity graph we will be using in this PhD thesis, but others like Beta-Skeleton [KR85] or Delaunay's triangulation [Del+34] are also very used.

This type of mathematical construct is suitable for structuring and modelling all kinds of phenomena or problems based on relational data. Vertices typically represent different elements, and edges typically represent the relationships between these elements. Depending on the particular field we are studying, we should choose a different type of graph. Directed edges can be helpful if the relationship between two elements is not symmetrical, like social graphs obtained from social networks, where the relationship can be mutual or one-sided. Positioned nodes can be helpful for modelling a geometrical problem in 3D space involving distances. Weighted edges can be helpful if these edges represent signals with varying power between electronic devices. This is the reason why graph theory can be applied to a wide variety of problems or models in academia. And not only in computer science, since it is also widely used in health sciences, linguistics, etc.

The main objective of this PhD thesis is to contribute to the fertility of graph theory, by applying it to different real-world problems. In particular, it covers the application of graph theory to the fields of wireless networks and opinion analysis. As we will be reviewing in further sections, this thesis also includes other research lines like electronic voting. In that particular case, we do not build upon results from graph theory, but from other mathematical areas, such as finite fields [MP13] or cryptography.

This thesis follows a compendium format, and therefore it reproduces in Chapters 2 to 6 the five articles published in peer-reviewed journals which compose my PhD work, [Gim+23; Gim+22; Tej+21; Ord+21; Lah+22]. At the beginning of each of those chapters, a brief discussion of the corresponding article is provided. Prior to that, in the following sections, I will give more details about how graph theory is applied to opinion analysis and wireless networks. In Section 1.3, I will explain the other research lines considered in this PhD. Finally, Chapter 7 summarizes the conclusions of all these works and sets future research lines.

1.1 Graphs applied to wireless networks

Networks are an archetypical example of application for graph theory, and computer networks are not an exception. In this research line we have worked with IEEE 802.11 networks, an specification widely known as Wi-Fi. These networks can operate in two modes: ad-hoc and infrastructure, being the latter the typical operation mode. In infrastructure mode, there are two kind of devices: access points and clients. Clients must connect to an access point in order to communicate with any other part of the network. All the clients attached to the same access point are communicating in the same frequency, but they are coordinated internally so that no internal interferences arise. However, each access point and its clients act independently from the others, acting on their behalf. Since all Wi-Fi frequency bands, on 2.4GHz, 5GHz or 6GHz, are limited in number of channels, such a behaviour often leads to situations where two nearby access points are using the same frequency channels. In consequence, interferences in the same frequency may exist between two access points and their attached clients. This, management of frequency channels is a critical field of research about wireless networks.

Wi-Fi networks can be represented as a weighted geometric graph. In Figure 1.1, we show an example or a residential building. Vertices represent devices, both access points and clients. In addition to a three-dimensional position, vertices also have an attribute indicating if they represent a client or an access point. Edges represent connections between devices, which can be measured to represent the intensity of the signal between a client and its access point, or the interference between a device and any other device communicating in the same frequency. For clarity purposes, interference edges are not displayed in the mentioned figure.

This graph model can be coupled with a propagation model and applied to a particular version of the Wi-Fi specification to simulate how Wi-Fi networks behave

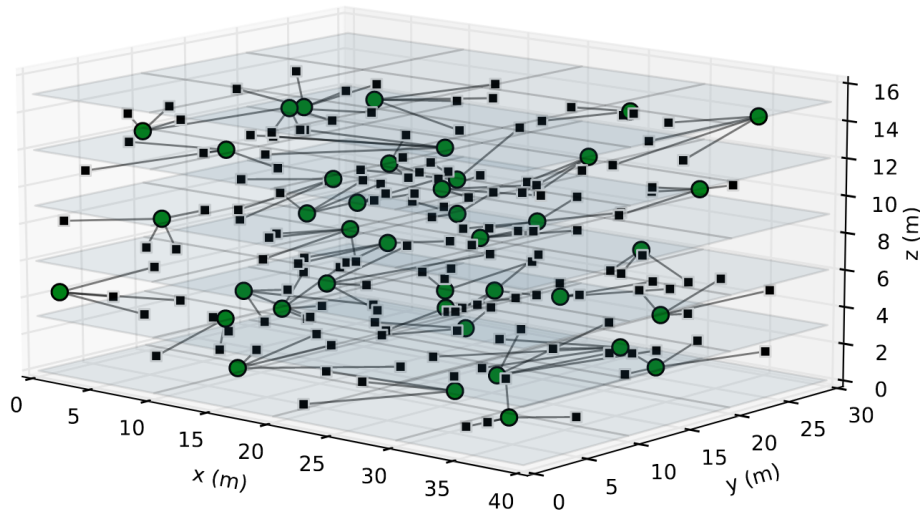


Fig. 1.1: Wi-Fi network of a residential building modelled in a graph, without interference edges. Extracted from [Gim+23]

in real life scenarios. The insights of our IEEE 802.11 graph model were published in Alexandria Engineering Journal [Gim+23] and can be read in Chapter 2. One of its exclusive features is the calculation of the co-channel interference. In some wireless frequency bands like in Wi-Fi 4 (2.4GHz), channels are partially overlapping. This characteristic makes the simulation of Wi-Fi 4 networks challenging. Our graph-based model offers a realistic simulation of the network by calculating different co-channel interference factors between vertices, adjusted by the channels selected by each access point. This model has been successfully validated in a set of realistic scenarios. As a result, we proposed a flexible yet powerful simulation model which allows us to research about the interaction between different parties sharing the same frequency spectrum.

With the help of this model, our motivation is to improve Wi-Fi networks performance. Previously, the cost of simulating a complete Wi-Fi network using a fully-featured network event simulator would make these research lines unfeasible. In addition, the current standard network simulators do not calculate co-channel interference, making the following research line impossible. In the article [Gim+22] reviewed in Chapter 3, we studied the technique of channel bonding in Wi-Fi 4. Channel bonding consists on merging pairs of adjacent frequency channels into a wider one. Wider frequency channels offer higher throughput of data transmitted, but, in exchange, they reduce the number of channels available. Channel bonding had been thought to be counterproductive when the number of channels available in the Wi-Fi 4 spectrum is 11. But we found a positive effect when using channel

bonding in countries where the Wi-Fi frequency spectrum includes 13 channels. The reason behind this result is the number of orthogonal channels available in each country. In the previous paragraph, we mentioned that frequency channels are partially overlapping. Two frequency channels are orthogonal if they do not collide. In countries where the 2.4GHz frequency band allows at least 13 channels, the number of orthogonal channels available when using channel bonding makes it beneficial.

1.2 Graphs applied to opinion analysis

Opinion analysis or sensory analysis has been studied in the past mainly under a statistical framework. For instance, *Napping* [Pag05] uses projective mapping, a technique for data collection which requires all participants to place samples in a two-dimensional space depending on how similar they are (similar samples should be close to each other, different samples should be far apart), and then uses Multiple Factor Analysis (MFA) [Pag05] to study the data obtained from all the participants. While accurate, MFA does not scale well with the number of participants in the study, since its complexity is cubic in that number.

Our research line starts with *SensoGraph*, an alternative method to analyze data from projective mapping using graph theory [Ord+19]. In *SensoGraph*, the input of each participant can be converted into a geometric graph using proximity graphs. Each sample will be a vertex with a position in two dimensions, while edges are defined in different ways depending on the specific mode we are using. We developed an efficient web implementation of the method [TO; OT] and exploited the fact that the complexity of this method is linear in the number of participants, instead of the cubic complexity of MFA [Ord+21].

One of the tools we use is the notion of proximity graphs, a particular type of graph which defines whether two vertices are connected or not depending on the neighbourhood condition. Each different type of proximity graph is defined by a different neighbourhood condition. The Gabriel Graph [GS69] is the proximity graph which worked best in *SensoGraph*. In this particular proximity graph, two nodes are linked if, and only if, the circle whose diameter is the potential edge between those vertices does not contain any other vertex inside. This step is called clustering in *SensoGraph*.

The original design used Gabriel Graph for the clustering step, but this step can be changed. Clustering is the process of converting a positioning of points into a

graph, following a determined algorithm. In the article [Ord+21], reviewed in Chapter 4, one of my contributions was designing a new clustering mode called ‘Distances’, which uses weighted graphs instead of proximity graphs in order to capture the essence of projective mapping. For this second clustering mode, all vertices are connected to each other, resulting in the complete graph. Then, the edges are assigned a normalized weight depending on their length, representing the similarity between the vertices they connect. The shortest edge (most similar pair) is weighted 1. The longest edge (least similar pair) is weighted 0. The rest of the edges are assigned a similarity weight using a spline curve which emphasizes extreme values aiming to highlight differences in similarity. In Figure 1.2, we can see an example of a weighted geometric graph and its adjusted adjacency matrix. The following expressions describe how similarity is computed:

$$s_{ij} = 1 - \frac{l_{ij} - l_{\min}}{l_{\max} - l_{\min}} \quad ; \quad s_{t,ij} = \begin{cases} 2^{p-1} \cdot s_{ij}^p & , \text{ if } s < 1/2 \\ 1 - 2^{p-1} \cdot |s_{ij} - 1|^p & , \text{ if } s \geq 1/2 \end{cases}$$

s_{ij} : Similarity between i and j , $s_{t,ij}$: Tuned similarity between i and j ,
 l_{ij} : Length of the edge (i, j) , l_{\min} : Length of the shortest edge,
 l_{\max} : Length of the longest edge, p : Highlight parameter

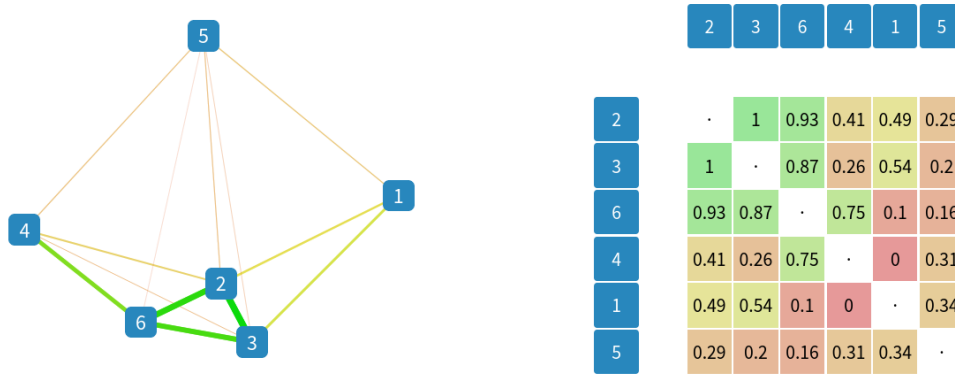


Fig. 1.2: Adjacency matrix which shows all edge weights in the ‘distances’ clustering mode.

In the same article [Ord+21], we evaluated both clustering modes of *SensoGraph*, using Gabriel Graph or measuring distances. We were able to produce accurate results using one of the largest datasets ever considered in projective mapping, leveraging the efficiency of *SensoGraph*.

Following this line, we created and tested a new method, called Linking, published in [Lah+22] (Chapter 5). Free-linking is a new input-based technique alternative to projective mapping. It requires participants to define a graph manually, that is, specifying edge existence one by one. This idea allows finer tuning than the input method free-sorting, which sorts products together in groups. Similarity is represented in a graph where all possible edges (all possible pair of samples) may exist or not independently from each other. In this article, we compared free-sorting with free-linking using a dataset of edges and show the results of this opinion analysis session.

The data was processed with DISTATIS [LAH18], a statistical framework which is the current standard for free-sorting. However, it can also be processed using the *SensoGraph* method, extracting adjacency matrices from the graphs.

Compared with free-sorting, its direct counterpart, and using the same tool (DISTATIS), we showed that the free-linking input method allowed a more precise and fine-grained analysis.

1.3 Electronic voting

Electronic voting (or e-voting) is a concept inside the field of electronic democracy, which consists on deploying a voting process with the same virtues offered by current physical ballot systems but automatized by electronic means, making the process less expensive in time and money. Some of the requisites for deploying electronic voting processes are already fulfilled, such as reliable networks or affordable powerful personal devices. But it is also essential to have a secure system which ensures the privacy and the integrity of the votes, among other basic properties.

The state of the art already had several secure protocols, which complied with the desirable properties for e-voting systems:

- **Ballot secrecy:** Only the voter knows the content of the actual vote.
- **Legitimacy:** Only authorized participants can vote.
- **Eligibility:** Legitimate voters can emit only one vote.
- **Accuracy:** The final result accounts for all legitimate votes, without any alteration.
- **Individual verifiability:** The voter can verify the correctness of the results.

- **Coercion resistance:** The voter cannot demonstrate the content of her vote to a third-party. In consequence, a malicious third party cannot be completely sure that the coerced person voted to the option she was told.
- **Robustness:** The system remains dependable even when some actors actively attempt to corrupt the process.

However, none of the systems in the literature guaranteed security for the voters without trusting the voting servers. This is why we developed the DiverSEC system, whose main motivation is to offer a secure e-voting system for both voters and voting authorities, where voters can detect any attempt of power abuse and authorities can detect any attempt of attack, making no assumptions about the correct behaviour of any agent. For this purpose, the vote will be fragmented and shared between voting authorities. These voting authorities are the servers of the candidates themselves, one server per candidate. This topology creates a power equilibrium.

Our system is based on three pillars: Shamir's secret sharing [Sha79], finite fields [MP13] and mixnets [Cha81].

The voting servers create a directed cycle graph, that is, these candidates define a ring, which is a kind of graph suitable for a cyclic mixnet. This is a common technique used in information security, and it works by aggregating multiple sources of data. All these messages are grouped together and then sent through several chained proxies, which shuffle the content they received from the last step. At the end, the receiver will obtain all the data without being able to trace the message to the source.

Votes are fragmented and shared thanks to a Secret Sharing Scheme. In this case, Shamir's Secret Sharing scheme in a finite field [Sha79]. Secret sharing is a mathematical cryptography technique which aims to split a secret in several parts under the following conditions. Being k the threshold, n the number of shares and s the secret:

- Knowing $a < k$ fragments does not reveal any information about s .
- Knowing $a \geq k$ fragments is enough to determine s unequivocally.
- All fragments are equivalent and interchangeable.

Shamir proposed an algorithm based on polynomials and 2D geometry. It is widely known that a line is determined by two of its points and a parabola is determined by three of its points. In other words, k points are enough to unequivocally determine a polynomial if its degree is lower than k . Translating this idea to the secret sharing

scheme, the secret s can be codified into the polynomial, the degree determines the threshold k , and the shares are all the points of the polynomial we want to distribute. For this scheme to be completely secure, we need to work in finite field (containing a finite number of elements). Typically, this is the equivalent of working with integers modulo a prime number. The consequence is that polynomials are not continuous curves anymore but discrete curves with a limited number of values. In Figure 1.3, a vote is divided into four shares, which are distributed along the voting servers ring.

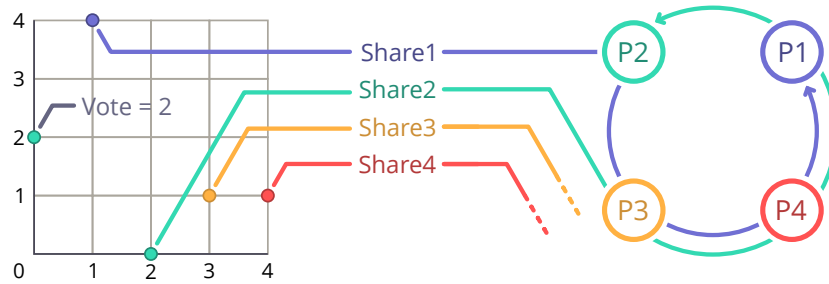


Fig. 1.3: Information flow of the vote shares in the cyclic graph topology

Relying on these three pillars we have achieved a secure e-voting system which is simple and reliable, and offers a friendly verification mechanism for voters. The system, described in Chapter 6, complies with the basic properties of e-voting systems except coercion resistance, which is the most difficult property to achieve in the current literature in remote e-voting systems.

IEEE 802.11 graph models

” *Eight-Oh-Two-Eleven. I suspect that’s where Eliza Cassan broadcasts from.*

— **Frank Pritchard**
(Deus Ex: Human Revolution)

A popular research line about 802.11 networks, commercially known as Wi-Fi networks, explores optimizing their performance through the optimization of their frequency channel distribution. These techniques and experiments require evaluating the results emulating a wireless network. In [Gim+23], we propose a graph-based model for Wi-Fi networks, which is specifically designed for this task, since it addresses co-channel interference. It is lightweight and also flexible, enabling research in different scenarios and different 802.11 specifications, such as Wi-Fi 4, 5 or 6.

The structure of the model is graph-based, since it fits perfectly the structure of a network. In particular, we have used geometric undirected weighted graphs. Vertices represent devices, both access points and clients, which have a position in the 3D space. Edges represent connections between two devices. Edges are bidirectional, that is why the graph is undirected. Also, each connection has several attributes, such as the power of the signal or the frequency channel.

The resulting model is both realistic and flexible, allowing researchers to simulate Wi-Fi networks without resorting to a heavy discrete-event simulator. The experiments described in the article take into account important parameters of a network, like client density, co-channel interference and 3D indoor propagation settings. For this reason, the model does not sacrifice accuracy. All these claims are validated in several experiments, from the most simple graphs to a set of generated realistic scenarios. This set of scenarios, covering a range of network densities, is openly available to the research community for benchmarking purposes. These validations include a comparison with the well-known network simulator *ns-3*.



Alexandria University
Alexandria Engineering Journal

www.elsevier.com/locate/aej
www.sciencedirect.com



ORIGINAL ARTICLE

IEEE 802.11 graph models



Jose Manuel Gimenez-Guzman ^{a,*}, Ivan Marsa-Maestre ^b, Luis Cruz-Piris ^b,
David Orden ^c, Marino Tejedor-Romero ^c

^a *Universitat Politècnica de València, Departamento de Comunicaciones, Spain*

^b *Universidad de Alcalá, Computer Engineering Department, Spain*

^c *Universidad de Alcalá, Department of Physics and Mathematics, Spain*

Received 29 July 2022; revised 13 October 2022; accepted 6 December 2022

Available online 14 December 2022

KEYWORDS

Wi-Fi networks;
Graphs;
Network modeling

Abstract There are several recent research lines addressing Wi-Fi network planning and optimization, both in terms of channel assignment and access point deployment. The problem with these works is that evaluation is usually performed with specific and closed models regarding signal propagation, throughput computation, and utility definition. Also, many of the models in the literature make assumptions about the role of wireless stations, or the co-channel interference, which -while being valid in the context of a single research work- makes very difficult to compare different approaches, to re-use concepts from previous mechanisms to create new ones, or to generalize mechanisms to other scenarios. This makes the different research lines in Wi-Fi network planning and optimization progress in an isolated manner.

This paper aims to address such a recurring problem by proposing a graph-based generic model for Wi-Fi utility computation in network planning scenarios, as well as providing a collection of scenario graphs which may be used to benchmark different planning and optimization approaches. Experiments are conducted to show the validity of the model and the significance of its features, along with its extensibility to other scenarios.

© 2022 THE AUTHORS. Published by Elsevier BV on behalf of Faculty of Engineering, Alexandria University. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Coexistence of multiple wireless networks and users competing for the scarce resources of the radio electric spectrum is a com-

plex problem demanding attention from the research community. Independent evolution and management of these networks have yielded an undesirable situation: most wireless networks are highly inefficient in many cases [1]. This is being addressed by the researchers from two different perspectives. On one hand, devising new standards and specifications for high-efficiency wireless local area networks (HEWs) [1]. On the other hand, improving planning, coordination, and optimization mechanisms for the existing standards.

In the latter case, a variety of approaches have been proposed, yielding promising results in a wide range of scenarios [2–6]. However, most of these approaches are evaluated with

* Corresponding author.

E-mail addresses: jmgimenez@upv.es (J.M. Gimenez-Guzman), ivan.marsa@uah.es (I. Marsa-Maestre), luis.cruz@uah.es (L. Cruz-Piris), david.orden@uah.es (D. Orden), marino.tejedor@uah.es (M. Tejedor-Romero).

☆ Peer review under responsibility of Faculty of Engineering, Alexandria University.

specific models not easily portable to other settings. Therefore, a recurrent challenge researchers have to face is how to justify the models and mechanisms they propose, or how to compare their approaches and methods with the ones of other researchers. In the best cases, there are a few number of previous works similar enough to the new proposal to make a comparison. In most cases, however, this comparison is not possible due to the diversity of the models the different research groups deal with, so the different research lines progress in an isolated manner.

This paper intends to bridge this gap by proposing a generic model suitable to be applied to different problems of Wi-Fi planning, coordination, and optimization, in order to enable comparison between different approaches in a usable and reproducible way. Our choice of a model is not arbitrary, and is motivated by the advantages and disadvantages of the different possibilities. Apart from using prototypes or real networks, which are expensive and difficult to set up, researchers usually resort to analytic and simulation models to design and evaluate the performance of telecommunication networks [7]. Each kind of approach operates from a different abstraction level, and therefore has its own advantages and drawbacks. While network simulators are able to capture a higher number of features, analytic models are faster and capture the essence of the feature under study, isolating it from the rest of features. Regarding drawbacks, analytic models are very limited by the underlying mathematical processes, while simulation models are computationally costly and usually not attractive to researchers due to the complexity of the underlying procedures (usually unknown to the researchers). Due to their advantages and disadvantages, the use of analytic and simulation methods coexist and complement each other. In fact, it is usual to initiate the studies with analytic models and evaluate them later via simulation before their implementation or commercial exploitation.

The proposed model is suitable for a wide range of wireless networks settings. However, we have focused our attention on those layouts where the high density of wireless devices makes coordination and optimization even more necessary to be able to obtain a satisfactory quality of service. These high-density settings are usually named dense or ultra-dense Wi-Fi networks and are being thoroughly studied by the scientific community [8,9] because, if not properly designed and optimized, they can become highly inefficient. In fact, the proposed model includes the paradigmatic issues that drive dense Wi-Fi networks to be inefficient. For example, dense Wi-Fi network models must be three-dimensional, as the interferences to a signal may come from any directions. Moreover, they must consider co-channel interferences, i.e. interferences from overlapping channels, as they can become very harmful and, therefore, they must not be ignored. Additionally, we consider that dense settings occur specially in indoor environments, so our model includes a realistic indoor propagation and loss model which considers obstacles. However, although we have tailored the model to include the most challenging features in Wi-Fi networks, the high expressiveness and flexibility of the model makes it also suited to be used for simpler settings and also to be extended to other situations. For example, to consider an outdoor layout we would only need to change the propagation and loss model. Or if, as in many previous works [10], we want to restrict the study to orthogonal Wi-Fi channels, we can directly use those and there will not be any kind of co-channel interferences. Finally, the expressiveness

of the model makes it useful to be the basis for more specific studies, like the consideration of channel bonding [11].

This paper proposes, to the best of our knowledge, the first generic model for Wi-Fi settings including realistic indoor signal propagation for three-dimensional settings, taking into account the precise location and interferences between all wireless devices (both APs and STAs), and considering co-channel interference among all available channels in Wi-Fi. The model makes contributions in four lines:

- We describe a model for IEEE 802.11 wireless communications, including architecture, signal propagation and interference, and throughput computation. The model integrates data from different sources, including the ITU propagation model, studies about indoor propagation and co-channel interferences and data about the relationship between Signal-to-Interference-plus-Noise Ratio (SINR) and actual throughput.
- We propose a graph-based modeling of realistic, three-dimensional scenarios, along with a collection of settings we put at the disposal of the research community. These settings represent realistic scenarios of dense Wi-Fi networks in residential buildings of multiple floors. These graphs can be used by the community for benchmarking, so that they can compare different approaches (e.g. their own against the ones in the state-of-the-art) in the same conditions.
- We conduct a comprehensive set of validation experiments, first with simple graphs that allow to check the validity and consistency of the model and then with the generated realistic settings, to assess the significance of the model properties. We can see there is a significant effect of the inclusion of vertical distance attenuation, co-channel interference, STA density and tri-dimensional layout, which makes this model more accurate and realistic than the ones used in previous works.
- We present two examples on how the model can be easily extended to accommodate other technologies. In particular, we adapt the model to channel bonding in Wi-Fi 5, and also to the study of the impact of the coexistence with Bluetooth devices.

The rest of this paper is organized as follows. The next section gives context to our proposal and frames it in the state of the art. Then, Section 3 describes our model for IEEE 802.11 wireless communication. Afterwards, Section 4 presents the scenario modeling based on graphs. Once the communication model and scenarios have been established, the experimental evaluation is described in Section 5. Finally, Section 6 briefly discusses our contributions and identifies lines for future research. As Supplementary Material, we provide the graphs for the generated scenarios, and we make the code for throughput computation using our model available upon request.

2. Related Work

There are a number of analytical and simulation models for wireless networks in the literature, some of which we have worked with in the past [12,13]. However, the diversity and high specificity of the models proposals for channel selection in WLANs makes very difficult to compare approaches [2].

Graph-based modeling has been widely used to model communication networks [14,15], specially for wireless networks [16–19]. However, their use for WLAN networks has been more limited [20,21]. Graph models in Wi-Fi (or wireless communications in general) have been used for frequency assignment problems (FAP) [22], specially when the problem has been studied as an instance of a graph coloring problem [16,23].

In the following, we describe some of the most prominent works that use graph models in the context of Wi-Fi networks. In [24], authors deal with a channel assignment problem considering interferences between APs, and they propose a weighted variant of the well-known graph coloring problem with realistic channel interference. Although [25] considers STAs in the channel assignment procedure, the proposed model does not include STAs as vertices in the graph, but it includes their effect by adding interference edges between APs when any of their associated STAs interfere. Similarly, in [4] authors propose a centralized channel assignment based on clusters to minimize the interference level at APs. In [26], we studied the channel assignment problem in Wi-Fi planar scenarios analyzing the effect of using only orthogonal Wi-Fi channels. Authors in [27] use machine learning for channel allocation in Wi-Fi networks using a simple graph model. In their model, vertices represent APs and edges between them exist provided both APs detect each other when scanning the spectrum. In [28], Stojanova et al. propose a graph model for Wi-Fi networks to measure performance using a Divide-and-Conquer strategy to break down the original problem into several sub-problems, and later combining the solutions to solve the whole original problem. In [10], authors propose both exact and heuristic wireless channel assignment techniques in the field of hybrid data center networks, thus enabling high throughput in wireless and wired transmissions. In this graph model, vertices represent communications between different transmitters and edges represent interference signals. The work described in [29] models a wide variety of Wi-Fi networks by means of graphs, and then computes some of their main features like selected centrality metrics. In [30], authors propose an enriched version of a conflict graph to model Wi-Fi networks, so it represents the partial conflicts (interference, detections...) between the devices in the Wi-Fi network, but it does not include STAs. In a similar way, [31] uses two conflict graphs (a physical one and a logical one) to study channel bonding in Wi-Fi networks, in. The vertices of the physical conflict graph represent APs and the edges represent the fact that two APs are neighbors, while the logical conflict graph depends on the channel assignment and, therefore, on the logical neighbors.

Once we have described the most outstanding or recent approaches in the field of graph modeling for Wi-Fi networks, we now describe and justify the need of our proposed graph model. Although the graph models we have found in the literature are valid for the purposes of each specific paper where they appear, they have different limitations which make them unsuitable for generic use and comparison among different researchers. Many of them use graphs that make unrealistic simplifications on the Wi-Fi propagation model, such as assuming on-off signal reception, that is, two wireless devices being either “in range”, and therefore fully receiving each other’s signal, or “out of range”, and therefore not interfering at all [24]. Those which use a more realistic propagation model,

such as [10,29], restrict the analysis to a limited set of available channels (3 or 4), hence ignoring co-channel interference, which is one of the most prominent aspects of Wi-Fi networks, as we have discussed in our previous work [32]. Moreover, our model is, up to our knowledge, the first graph-based model that uses a realistic indoor propagation model like the one described by ITU in Recommendation P.1238 [33]. On the other hand, although it is a wide belief in the community that Wi-Fi assignments using only orthogonal channels yield better results, this has only been studied for settings with planar scenarios [26]. However, most of the realistic settings where Wi-Fi networks are deployed, such as residential buildings or offices, are three-dimensional, where wireless devices in different floors may interfere with each other. Finally, models in the literature usually do not take into account the wireless stations (STAs), considering only the access points (APs) [4,27,28,30,31], or modeling the effect of STAs in a simplified manner (e.g., by assuming that two APs also interfere if two of their clients are in range of each other, but without taking into account signal propagation between them) [34].

Taking all this into account, we have developed a streamlined, flexible and reusable yet accurate generic Wi-Fi model framework that can be used by other researchers to propose, implement, evaluate and compare their work with other state of the art proposals. In addition, our framework lays somewhat in the line between analytic and simulation models. We discarded the use of pure analytical models (such as the one described by a queueing network or a Markov process) because the complexity of Wi-Fi networks as a whole makes it very difficult to analytically solve them. In the same sense, we discarded to use a discrete event simulator for Wi-Fi due to its high computational cost and the fact that the high number of parameters that a simulator requires makes the scientific proposals not easily replicated. Finally, most commercial simulators do not consider adjacent channel interferences [35], which are specially critical in dense settings. So we propose a model which is closer to simulation than the usual analytical models, but that is usable and flexible enough to be used for comparison in the earlier and middle stages of research, prior to implementation, where simulators are definitely needed for final tuning.

3. Modeling IEEE 802.11 wireless communications

3.1. IEEE 802.11 architecture

Wi-Fi networks are composed of two different devices: access points (APs) and stations (STAs). Although Wi-Fi networks can also operate in *ad hoc mode*, we focus in the most widely used operation mode, which is *infrastructure mode*. In this mode, communication always occurs between an AP and a STA, since direct communication between STAs is not permitted. Therefore, to provide connectivity to STAs, they must be associated to an AP. For that reason, in this operation mode, Wi-Fi networks can be seen as a set of *clusters*, where each cluster is an organizational structure made up of an AP and all its associated STAs. Fig. 1 shows an example of a Wi-Fi network operating in infrastructure mode composed of 16 APs (large green circles) and 54 STAs (small black squares), where the concept of Wi-Fi cluster can be clearly observed.

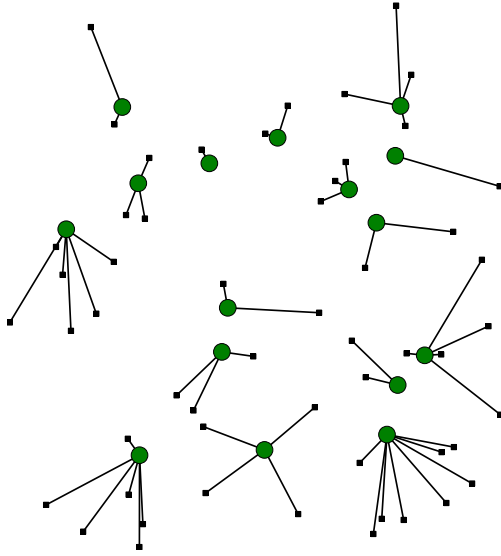


Fig. 1 Example of Wi-Fi deployment using infrastructure mode.

Without loss of generality, in this paper we model the IEEE 802.11n standard (recently renamed as Wi-Fi 4) operating in the 2.4 GHz frequency band with 20 MHz channels. However, the model is not restricted to a specific technology, since it is an intermediate abstraction level between analytical models and discrete event simulations. For that reason, and for example, it would be straightforward to use the model for IEEE 802.11ax (Wi-Fi 6) by tuning the configuration parameters. The reason of choosing the 2.4 GHz frequency band is because it is the most congested of the unlicensed frequency bands, which makes it specially challenging for network planning and optimization.

3.2. Signal propagation and interferences

To describe how wireless signals propagate, we have used the model defined by the ITU-R in Recommendation P.1238–10 [33], which describes an indoor transmission loss model for APs and STAs in the same building. Note that this model also considers the losses produced when the signal traverses different building floors. In [33], propagation losses expressed in dB are defined as:

$$L_{total} = 20\log_{10}f - 28 + N\log_{10}d + L_f(n), \quad (1)$$

where f is the frequency in MHz, N is the distance power loss coefficient, d is the distance between transmitter and receiver in meters and $L_f(n)$ is the floor penetration factor when signal traverses n floors. As previously mentioned, we have focused our attention in the most widely used 2.4 GHz frequency band, as it is the most congested one. However, it would be straightforward to consider another frequency band like 5 GHz. For the frequency band considered here, [33] specifies a value of $N = 28$ for residential environments. However, it is also admitted in [33] that propagation through walls increases the power loss coefficient considerably, using as examples paths between rooms in closed-plan buildings. For that reason, and according to [36], we have used $N = 28$ when $d < 16$ meters and $N = 38$ for $d \geq 16$ meters. Moreover, we have considered $L_f(n) = 10n$, since [33] states that losses across two floors in residential envi-

ronments are 10 dB when using concrete. With these values, propagation losses in indoor environments only depend on the distance and the number of floors traversed.

Once wireless signal propagation losses have been defined, we can derive the signal power (in dBm) received by a wireless device i (AP or STA) from the signal transmitted by a wireless device j , as:

$$P_r^{j \rightarrow i} = P_t^j + G_j + G_i - L_{total}, \quad (2)$$

where P_t^j stands for the transmission power of Wi-Fi device j (expressed in dBm) and G_j and G_i represent the transmission and reception antenna gains, respectively, both expressed in dB.

Signals that are received by a wireless device can either be the desired signal or an interference (undesired) signal. In the latter case, the power of such an interference will also be affected by two other factors. First, we consider that, due to the carrier sense multiple access with collision avoidance (CSMA/CA) mechanism used in Wi-Fi, APs and STAs cannot transmit continuously. To account for that issue, some studies [37] have modeled that mechanism as a continuous time Markov chain (CTMC). As a result, when both STAs and APs want to transmit a packet with probability P_{STA} and P_{AP} , respectively, they will succeed with a certain probability P_s . Also, P_{AP} is assumed to be greater than P_{STA} , as the amount of data transmitted by APs is expected to be higher than that of the STAs. Both phenomena are included in our model by means of a factor ψ , considering that this factor is different for APs and STAs ($\psi_{STA} = P_{STA} \cdot P_s$ and $\psi_{AP} = P_{AP} \cdot P_s$). Second, we consider that interferences are not always in the same frequency channel, but they can be in any other channel whose spectral mask collides with the frequency channel that the receiver is using. This effect is one of the most prominent peculiarities of Wi-Fi networks, as frequency channels where wireless devices can operate are partially overlapped, and the interference is higher the closer the channels are in the spectrum. More specifically, we use the factor $\kappa(|c_j - c_i|)$ to represent the co-channel interference of two interfering wireless devices using channels c_j and c_i respectively, where $|c_j - c_i|$ is the “distance” (in frequency terms) between the channels. To account for this co-channel interference we have used the values defined in [38], as shown in Table 1.

Including these two factors in the received power of interference signals, we can compute the received interference power at device i due to the signal emitted by device j as (in linear scale):

$$\tilde{P}^{j \rightarrow i} = P_r^{j \rightarrow i} \cdot \psi \cdot \kappa(|c_j - c_i|). \quad (3)$$

3.3. Throughput computation

Once the power of the desired signal and received interferences is defined, we can compute the Signal-to-Interference-plus-Noise Ratio (SINR) in each device. To compute SINR experienced at a certain wireless device i whose desired signal comes from a device χ , we use the following ratio:

$$SINR_i = \frac{P_r^{\chi \rightarrow i}}{\sum_{j \in \mathcal{J}} \tilde{P}^{j \rightarrow i} + N}, \quad (4)$$

Table 1 Spectral overlap between Wi-Fi channels [38].

$ c_j - c_i $	0	1	2	3	4	5	≥ 6
$\kappa(c_j - c_i)$	1	0.8	0.5	0.2	0.1	0.001	0

where $P_r^{z \rightarrow i}$ is the received power of the desired signal, \mathcal{J} is the set of devices that emit interference signals and N is the power of the thermal noise. To compute N (expressed in dBm) we use the equation:

$$N = -174 + 10\log_{10}(\Delta f), \quad (5)$$

where Δf is the bandwidth of the channel used expressed in Hertz. As we consider the use of 20 MHz channels, we have a thermal noise of $N = -101$ dBm.

Wi-Fi networks use an adaptive modulation and coding scheme (MCS) that depends on the available SINR, having a set of predefined MCSs defined by an index. Then, depending on the MCS in use, we will have a certain throughput. Obviously, as the SINR is higher we will be able to use modulations with a higher number of bits per symbol and coding schemes with a higher proportion of useful information (or less redundant information), which will yield better throughputs.

As mentioned above, we have focused on Wi-Fi 4 [39] using 20 MHz channels. Within Wi-Fi 4, there is another parameter which is relevant to define the achievable throughput for each MCS: the Guard Interval (GI), that represents the pause between packet transmission. Wi-Fi 4 defines two possible Guard Intervals, 400 ns and 800 ns. Although the achievable throughput is higher with the shortest GI, we are focusing on $GI = 800$ ns, since it is the mandatory option that must be implemented and is widely used as the default value. Finally, provided a certain channel bandwidth and a GI, we have to define the different SINR thresholds that determine the use of a certain MCS. For that purpose, we have used the values from [40]. Table 2 shows the relation between MCS, SINR and throughput. It is important to highlight that STAs have two different throughput values: downlink and uplink. Downlink throughput is related to the information emitted by the AP that is sent to the STA, so it can be computed measuring at the STA the SINR of the signal from the AP to the STA. Analogously, uplink throughput is computed using the SINR of the signal from the STA to the AP measured

at the AP. Finally, we have also considered that if the desired signal power ($P_r^{z \rightarrow i}$) is below the sensitivity of the receiver (\mathcal{S}), the throughput is zero, so it will not be necessary to evaluate interference and noise power.

4. Model proposal

4.1. Graph model

Once we have described the models used for computing the throughput achieved by the different network elements composing the Wi-Fi network, we can now define specific layouts that represent realistic Wi-Fi deployments. To model Wi-Fi network layouts we use geometric undirected graphs.

A graph can be defined as a pair (V, E) , where V is a set of vertices, and E is a set of edges between the vertices $E \subseteq \{(u, v) | u, v \in V\}$. In our proposal there are two types of vertices, corresponding to the two types of wireless devices we can find in Wi-Fi infrastructure networks: access points (APs) and stations (STAs). We use geometric graphs, so that each vertex in the graph has a position corresponding to the location of its corresponding wireless device in the network deployment. In addition to having two types of vertices, we also have two types of edges, as edges can represent either the association between APs and STAs or the interference between wireless devices. Since our graph will only represent symmetrical information, like the distance between vertices, we use undirected graphs. However, actual interferences are not symmetrical, because of the factor ψ (see Section 3.2).

In the proposed graph, if two wireless devices are associated (one of them must be an AP and the other a STA) they will be linked by an edge of type *signal*. An AP and all its associated STAs belong to the same *signal cluster*. A vertex v will be linked by an edge of type *interference* to each of the vertices of the graph that belong to a different signal cluster.

Table 2 Relation between MCS, SINR and throughput in Wi-Fi 4 using 20 MHz channels with mandatory 800 ns GI [39].

MCS index	Modulation scheme	Coding rate	Throughput (Mbit/s)	SINR range (dB) [40]
0	BPSK	1/2	6.5	[6.8, 7.9)
1	QPSK	1/2	13.0	[7.9, 10.6)
2	QPSK	3/4	19.5	[10.6, 13.0)
3	16-QAM	1/2	26.0	[13.0, 17.0)
4	16-QAM	3/4	39.0	[17.0, 21.8)
5	64-QAM	2/3	52.0	[21.8, 24.7)
6	64-QAM	3/4	58.5	[24.7, 28.1)
7	64-QAM	5/6	65.0	≥ 28.1

One of our aims is to propose a Wi-Fi graph model which produces graphs that are easily usable by other researchers. To show this and to explain the internal structure of the proposed model, Fig. 2 shows two complementary representations of the graph model corresponding to an example deployment with only two APs and two STAs per AP. On the left, we show the association relationships between APs and STAs, and the properties related to each vertex element in the graph model. We can see that, in addition of the key that unambiguously identifies every vertex, vertices have the property *pos* to include their position in the space, defined by three coordinates. In addition, the property *floor* determines the floor of the building where the vertex is. Moreover, the property *type* determines if the vertex is an AP or a STA. Finally, for APs we have property *listSTA*, which is a list of the keys of all its associated STAs. On the other hand, for STAs we have property *associatedAP*, containing the key of the AP the STA is associated to. Fig. 2, on the right, shows the signal/interference information from the perspective of STA-3, and also the properties related to edges in the graph model. We can see that edges have the *dist* property, representing the distance between the two vertices connected by the edge. Moreover, property *type* determines if an edge represents an association AP-STA (then the value is *signal*) or if it is an *interference*. It is important to highlight that new properties can be easily added to vertices and/or edges after loading the graphs, to support the different applications or problems for which the model can be used. Examples of such properties could be the Wi-Fi channel where each vertex operates, which could be added to each vertex in the graph, or the power losses for each edge in the graph, which could be added to edges. This makes easy for researchers to use their available code for their coordination or optimization mechanisms in our model.

4.2. Wi-Fi scenarios modeled

In this subsection we describe the specific layouts of the graph models proposed. To download these graph models, please see the “Supplementary materials” section. The generated graphs represent a realistic setting of 5-floor residential buildings, a paradigmatic example of a dense Wi-Fi network. Each floor has a dimension of $40 \times 30 \times 3$ meters (length, width and height respectively) and has eight flats in a 4×2 arrangement.

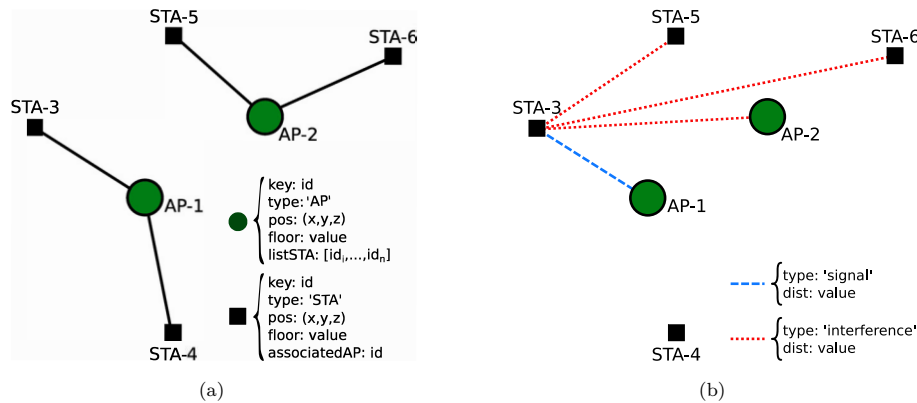


Fig. 2 Illustrative example of the Wi-Fi graph model and properties: (a) AP-STA associations and vertex properties; (b) signal/interference information from the perspective of STA-3 and edge properties.

As usual in residential layouts, in each flat there is a single AP. However, the number of STAs in each flat (η) varies. We have designed graphs with values of η ranging from 1 to 12 to represent a wide range of density of STAs. In all cases, every STA inside a flat is associated to the AP of the same flat. The position of both APs or STAs in the x- and y-axis is randomly generated from a uniform distribution bounded to the limits of the corresponding flat. Regarding the z-axis, each AP or STA is randomly located according to a normal distribution with mean 1.5 meters and a standard deviation of 0.5 meters, also bounded by the floor limits. To summarize, the number of APs in all graphs is equal to $8 \times 5 = 40$ APs, while the number of STAs ranges from 40 (when $\eta = 1$) to 480 (when $\eta = 12$). Finally, for each value of η we have generated 10 different settings, for a total of 120 scenarios.

For the sake of space, we only show a graphical representation of six of the total 120 scenarios generated. More specifically, Fig. 3 shows one sample scenario for several values of η , being η the number of STAs per AP. Fig. 3 only depicts the edges between vertices that represent the association between APs and STAs, avoiding, for the sake of clarity, the representation of interfering signals. To provide a more in-depth representation, we focus on the simplest scenario shown in Fig. 3, i.e., in the layout where $\eta = 2$. For that scenario, Fig. 4 shows the graph representation as projections to the different axes to be able to analyze the graph under different points of view.

Note that Figs. 3 and 4 do not show the interferences between Wi-Fi devices. To show the high number of interferences that appear even in the simplest scenarios ($\eta = 2$), we include two different representations in Fig. 5. Fig. 5a shows the interference edges (in red) produced by a single STA (in blue, with the edge to its associated AP also shown in blue). However, to have an idea of the whole graph including all associations and interferences, we can resort to Fig. 5b, where the very high density of interferences even for $\eta = 2$ is shown.

5. Experimental evaluation

In this section we validate the proposed Wi-Fi model under a variety of situations. First, we use a very simple graph to understand how the model operates and to check its consistency. Later, we make some experiments with the abovementioned

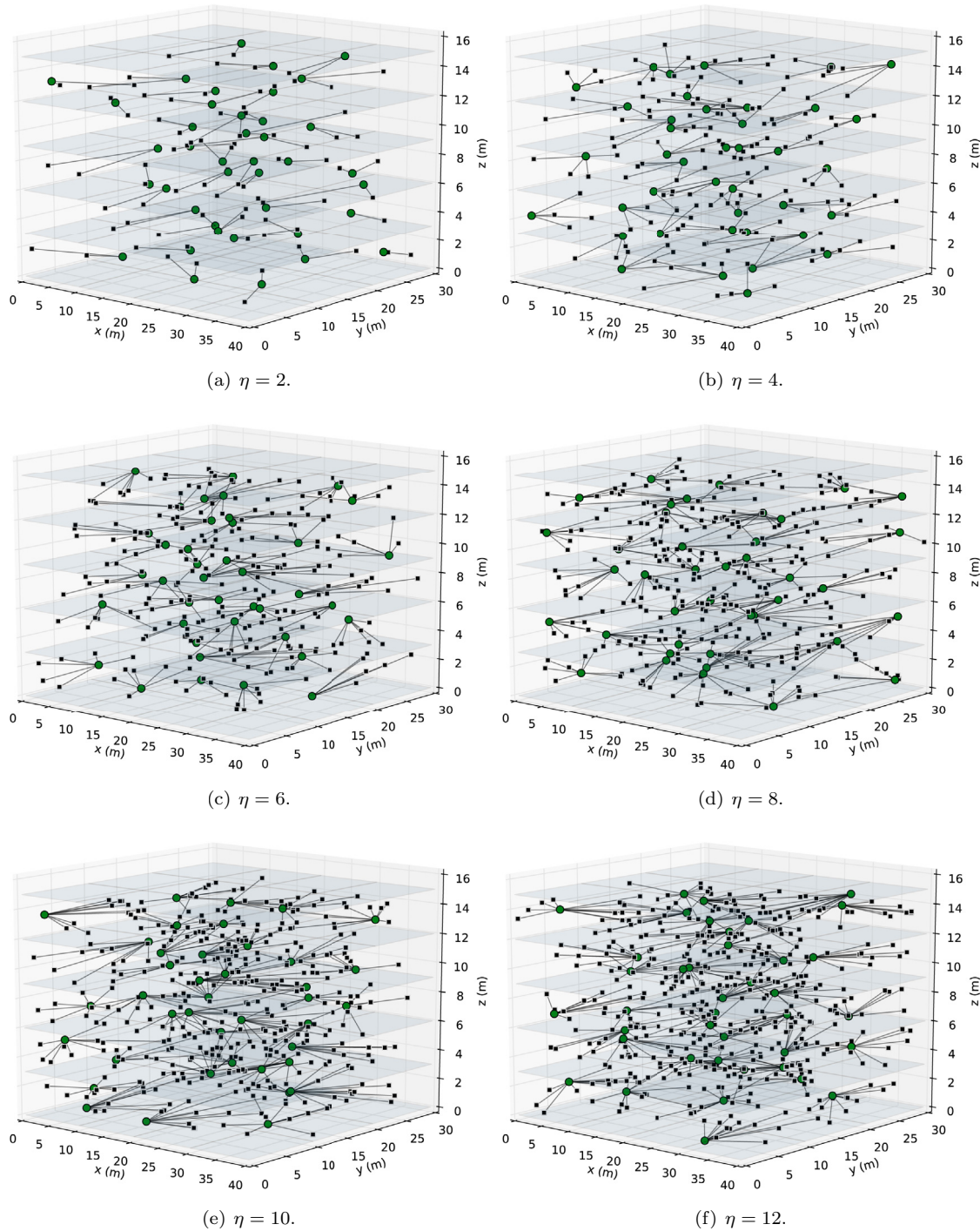


Fig. 3 Representation of some of the layouts modeled.

tioned 5-floor building graphs and evaluate the performance of the model as well as the significance of some of the features that it incorporates, such as the three-dimensional layout or the co-channel interference. Finally, we show the flexibility of the model by extending it to handle two additional different scenarios: channel bonding in Wi-Fi 5 and interferences due to coexistence with Bluetooth devices.

Our model, in order to be easily used by other researchers, does not include a high number of parameters to be set up.

However, there are some minimal ones that must be tuned. Although this configuration can be very easily changed by the user, in Table 3 we provide a reference for these parameters. Note that these reference values have been chosen for being either typical or reasonable values, and are the ones used in the sections that follow. The different parameters that are shown in Table 3 are the transmission power of wireless devices (P_t), the gain of the transmitting and receiving antennas (G_t and G_r respectively), the sensitivity of the receiver

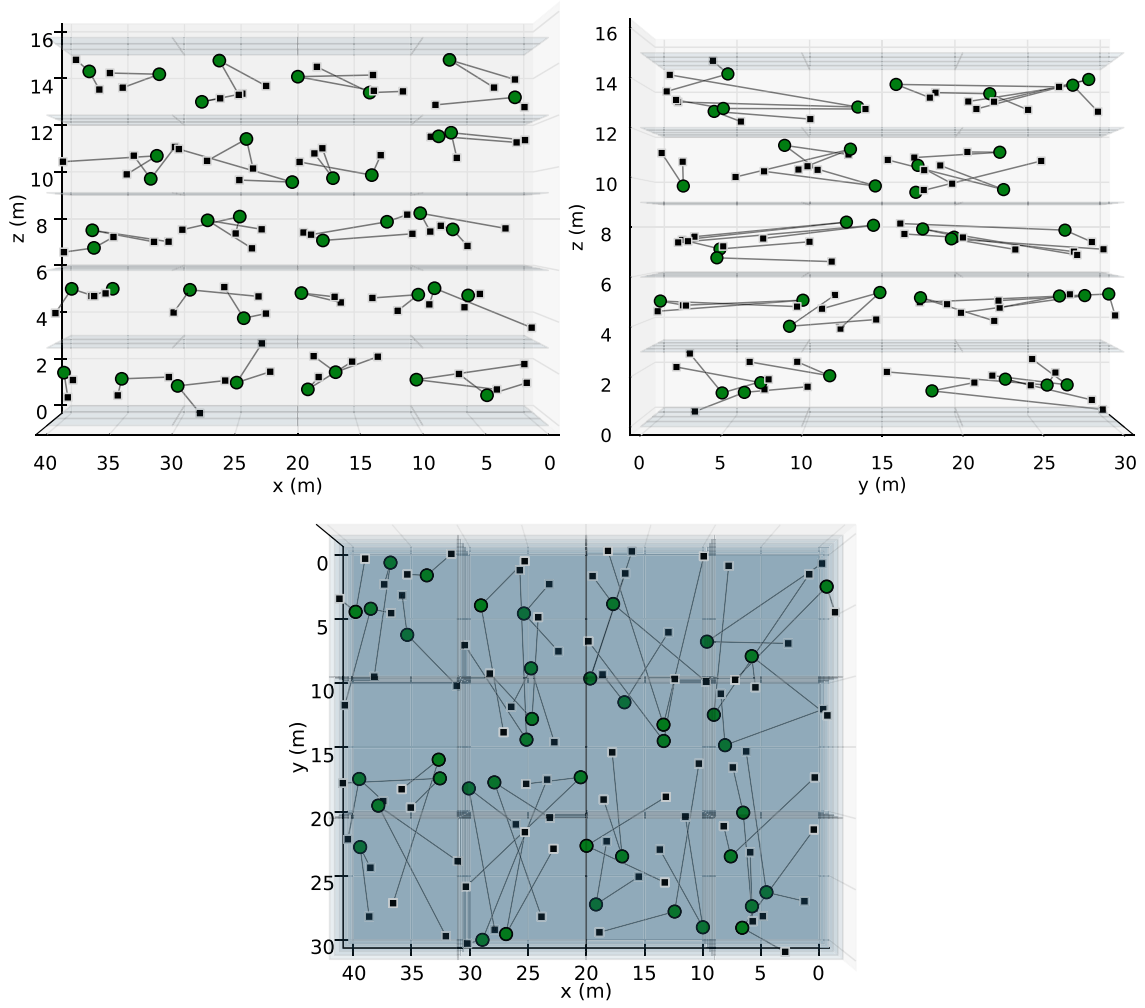


Fig. 4 Projections of a scenario with $\eta = 2$.

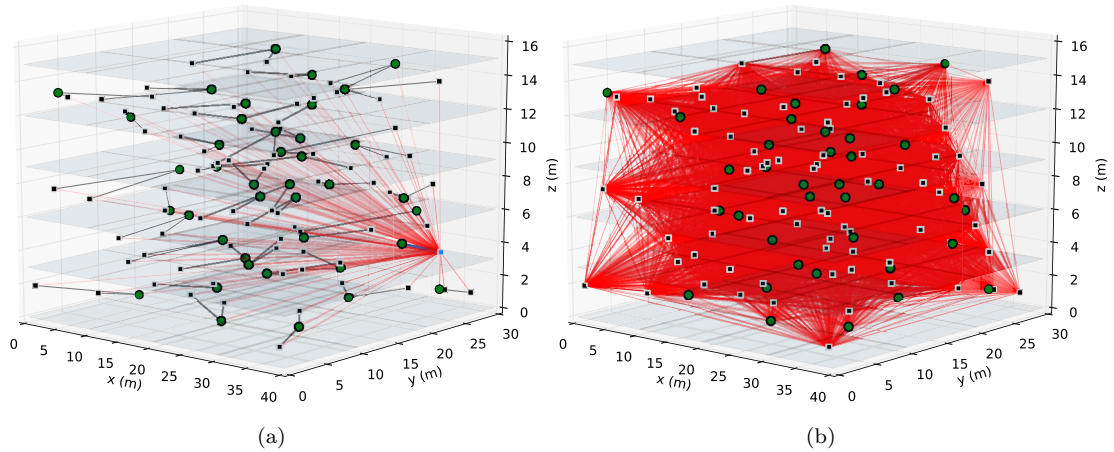


Fig. 5 Graph representation including interferences in a scenario with $\eta = 2$.

(\mathcal{S}) and Ψ , that models the transmit and success probability of APs and STAs.

To be easily reusable, one of the design requirements is to avoid very specific programming languages or libraries. To this end, our model relies on Python and the widely used *NetworkX*

library. Although the model can be extended, our experiments are conducted using 20 MHz channels in the most congested 2.4 GHz frequency band. Moreover, we have assumed that STAs attach to their corresponding AP, which is the typical situation in residential environments. Another choice could

Table 3 Summary of parameters.

Parameter	Value
P_t	30 mW
G_t	0 dB
G_r	0 dB
\mathcal{S}	-85 dBm
Ψ_{AP}	0.5
Ψ_{STA}	0.1

have been that each STA associates with the AP with the highest SINR, which is the usual situation in enterprise environments, where there are several APs for a single Wi-Fi network.

5.1. Validation of the model

To validate the model we have chosen a simple experiment, evaluating how the throughput and SINR change when a STA moves away from its associated AP. For this purpose, and to avoid any other effect, we have used a very simple graph that consists of a single AP with a single STA, so there are not interferences and we can focus in how the throughput changes with the distance. Additionally, we use this setting to validate the proposed model running the same experiment also with a well-known discrete event simulator: ns-3 [41]. Note that to be able to compare both proposals, we have used in ns-3 the same indoor propagation model as in our model: ITU-R P.1238-10. Moreover, we have setup the Wi-Fi manager in the simulator according to the settings used in our experiments. Finally, and regarding the traffic load of the different network elements, as our model computes the highest reachable throughput that STAs can obtain, in the simulator we have considered greedy traffic sources emitting UDP datagrams with a rate higher than the maximum throughput that can be obtained by Wi-Fi networks.

Fig. 6 shows how SINR and throughput change when the horizontal distance (in the same floor) between AP and STA changes, both using our model and the discrete event simulator. First of all, we can notice that the SINR obtained by our model and the simulator completely coincide. However,

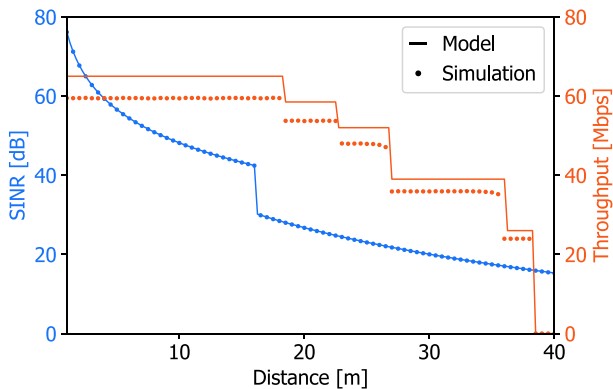


Fig. 6 Effect of the horizontal distance between AP and STA in throughput and SINR.

when we inspect the throughput, we note an offset between both models. The reason of this offset is because ns-3 measures the throughput at the application layer (usually called goodput) and our model measures it at the physical layer. Therefore, the difference between both curves is due to the overhead introduced by the different protocols operating in the different layers of the protocol stack, so we can conclude that our model, in comparison to a well-known discrete event simulator, is valid.

We now inspect the shapes of the curves themselves, as we have chosen a simple setting also to check if the behavior of the model matches the theoretical expectations. We can see that SINR exponentially decays when the STA moves away from the AP. Moreover, when the STA is 16 meters away from the AP there is an abrupt decay because of the propagation model used, which changes the power loss coefficient N at that distance (see Eq. 1). However, we can see that this discontinuity in SINR computation has no effect on throughput computation, since the STA is able to get the maximum available throughput (65 Mbps using MCS 7) until it is 18.4 meters away from the AP. Further from this point, it cannot use MCS 7 anymore and uses MCS 6 instead, decreasing the throughput down to 58.5 Mbps. A similar behavior can be observed when the distance is 22.62, 26.96, and 36.07 meters, as the MCS used changes to MCS 5, MCS 4, and MCS 3, obtaining 52, 39, and 26 Mbps, respectively. However, when the distance is 38.32 meters, there is an abrupt decay in throughput, that equals to zero from that distance on. The reason for this decay is that the power of received signal at the STA is lower than the sensitivity of the receiver ($\mathcal{S} = -85$ dBm), so it cannot be decoded. Note that in this case the sensitivity is limiting the use of MCS 2 to MCS 0. However, this does not mean that these MCS indices will not be used in any case, because there can be situations when the received signal is higher than \mathcal{S} but, due to undesired interferences, SINR is in the range of those MCS indexes.

5.2. Effect of vertical distance in throughput and SINR

After evaluating the impact on performance due to a horizontal movement of the STA, we can now analyze, in the same setting as in the previous section, the impact of moving the STA vertically, i.e., across different floors, so we can show the impact on performance of traversing different floors. To perform this experiment, we have situated the AP at ground level and moved the STA vertically, starting from the same position of the AP. Following this movement, the STA will change the floor every 3 meters, provided that this is the height of each floor. Fig. 7 shows the results of such vertical movement. We have used green dashed vertical lines to represent when the STA goes up to the next floor. As we move the first 3 meters, the behavior is the same as in the horizontal movement, as we are moving inside the first floor. Regarding the obtained SINR, when $d = 3$ meters we notice the first 10 dB decay in the SINR due to the effect of losses across adjacent floors. This behavior is repeated every time that the STA goes up a floor. Moreover, there is another abrupt decay at $d = 16$ meters, because of the change of the power loss coefficient from $N = 28$ to $N = 38$ (which again does not affect throughput computation). If we analyze the achieved throughput, we notice that it is possible to use MCS 7 (and therefore the max-

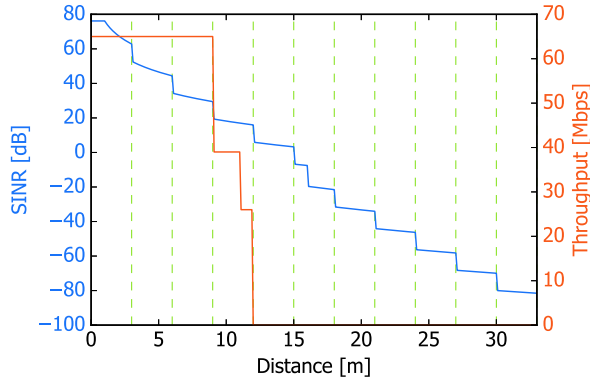


Fig. 7 Effect of the vertical distance between AP and STA in throughput and SINR. Green dashed vertical lines represent a change of floor.

imum throughput) even in the third floor. However, there is an abrupt decay in the throughput in the fourth floor, as we use MCS 4 when the STA is close to the floor and MCS 3 in positions of the fourth floor close to the ceiling. Finally, the received signal is below the sensitivity when we are in the fifth floor, so the STA cannot keep connected and the obtained throughput is zero. At this point, it is important to remember that this behavior is a maximum performance bound of the signal quality that could be obtained when we move across different floors, as in this experiment we are not considering other Wi-Fi devices that cause harmful interferences. Finally, if we compare Figs. 6 and 7, we notice that, as expected, the signal quality degrades much faster when we move vertically than it did when moving horizontally.

5.3. Effect of the density of STAs

Now, we evaluate the proposed model in terms of the throughput achieved when the density of STAs changes. For this evaluation we make use of the three-dimensional 5-floor building graph described in Section 4.2. In this model, there is a parameter, named η , that describes the number of STAs associated to each AP in the graph. This parameter ranges from $\eta = 1$ (there is a STA associated to each AP), until $\eta = 12$ (with 12 STAs per AP), so it yields experimental settings with a wide diversity of STA density.

However, before using the graph model to compute the throughput in this setting, we have to define how the different APs are configured in terms of which Wi-Fi channel they are using, since now we have to take into account the effect of interferences, and these vary greatly depending on channel assignments. Although there have been many specific proposals to assign channels to the different APs in Wi-Fi networks [2,24,10,42,43], in order to make the model as general as possible, we use a generic procedure inspired in [44], which is also the *de facto* standard in commercial, residential access points. This procedure is widely used and consists of each AP scanning the spectrum and using the frequency channel where it finds the least interfering power coming from other signals. Once we have assigned channels to all APs in the setting using this approach, we have computed the average uplink and downlink throughput for the 10 different scenarios we have for each value of η . In Fig. 8 we show the mean and 95% con-

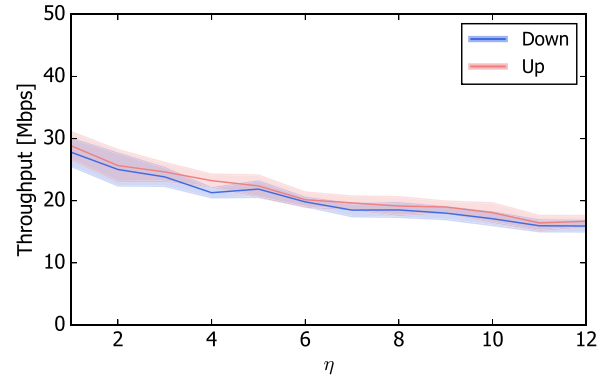


Fig. 8 Effect of the density in throughput.

fidence intervals for both uplink and downlink throughput for different values of the density η . We can see that there are not statistically significant differences between the uplink and downlink throughput values. Also, as expected, we notice that the throughput decreases as the density of the layout increases, due to the fact that there are more interferences in settings with higher densities.

Although we have evaluated the average throughput depending on the density of STAs η , in the following we are performing a more in-depth evaluation to gain insight on how a particular density value η affects the throughput achieved by specific STAs. For the sake of space, we limit the presentation of results to two specific layouts with different values of η , namely $\eta = 3$ and $\eta = 8$. Moreover, as the conclusions are equivalent for both uplink and downlink throughput, we only show the results for downlink traffic. Fig. 9 shows the downlink throughput obtained by each STA when $\eta = 3$ (Fig. 9a) and $\eta = 8$ (Fig. 9b). For each deployment, the figure shows the 5 different floors vertically, so the lower figure represents the first floor, while the upper figure represents the fifth floor. In each floor we also show the 8 different apartments in a 4×2 arrangement. Moreover, it is important to highlight that the layout has been represented with five layers (representing each layer a floor), but the position in the z-axis of each vertex in the graph is a continuous parameter. For the sake of clarity in the representation, we have projected all the STAs and APs in a floor to a plane. However, two Wi-Fi elements can be very close in the z-axis even when they are in different floors, which must be taken into account before interpreting the results, since discretization of the continuous range of heights implies that vertical distance information is partially lost (a single floor represents a whole range in the z-axis position). As expected, to achieve a high throughput, the distance from the AP to the STA plays a paramount role. Moreover, we notice low or even zero throughputs for STAs that are far from the AP, specially if there are many devices from other clusters in another apartment (of the same floor, but also from adjacent floors). We can also see that, in the high-density scenario, floors other than the first and fifth ones have significantly more STAs with the lowest values of throughput, and the middle floor is the one showing less STAs with the highest values of throughput. This is consistent with the expectations, since the center floor is the one which has higher interference from other floors. As expected, first and fifth floors have, on average, better throughput.

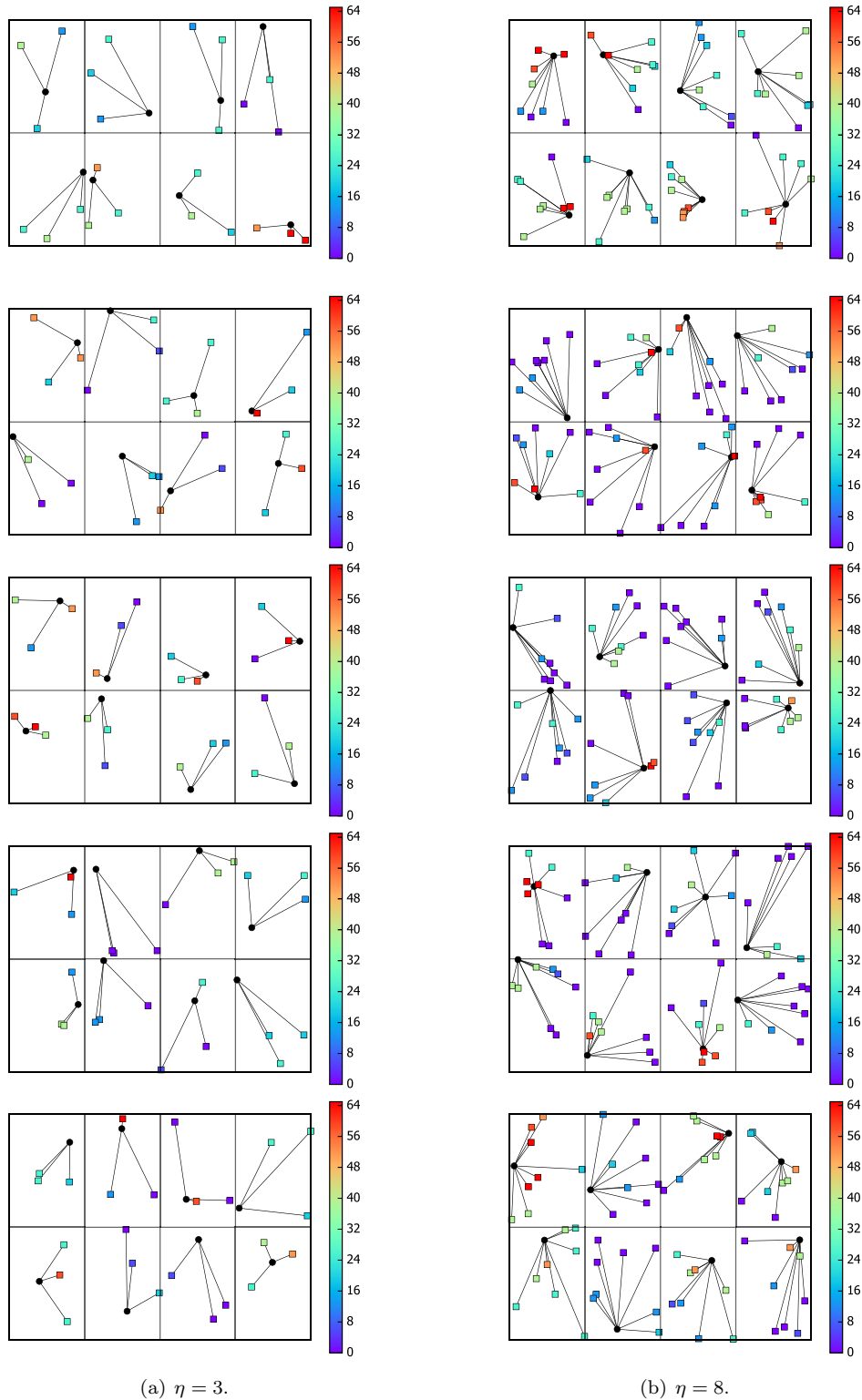


Fig. 9 Downlink throughput (in Mbps) in each STA. From down to top, floors 1 to 5.

5.4. Effect of the co-channel interference

The partial overlap between the frequency channels where Wi-Fi can operate is the main peculiarity of Wi-Fi networks operating in the 2.4 GHz band. As this peculiarity does not

appear in other communication networks, the co-channel interference is not usually considered in models, specially those models used by discrete event simulators. In this section, we compare the throughput that is obtained when we consider the co-channel interference (as our model does) with the per-

formance when we consider that Wi-Fi channels are not overlapped, to evaluate whether the effect of co-channel interference is negligible or not. Fig. 10 shows this effect, proving that the effect of the co-channel interference in the throughput is significant. When we do not consider co-channel interference (i.e., only devices using the same channel are assumed as undesired signals) interferences are much lower and thus the obtained throughput is much higher, but this does not represent the real behavior of Wi-Fi networks, where devices operating in “close” channels actually do interfere. In summary, co-channel interference has a significant impact on performance, so it is a crucial issue to include in a realistic Wi-Fi model.

5.5. Effect of using a three-dimensional layout

In the literature related to research in Wi-Fi networks it is usual to consider planar scenarios. However, in some scenarios, like residential ones, the use of three-dimensional (3D) settings should be considered. In this section we evaluate the importance of considering 3D-layouts instead of the classical 2D ones. For that purpose, we have evaluated the average and 95% confidence intervals of the downlink and uplink throughput for different values of STA density η . More specifically, to study the throughput avoiding the z-axis we have ignored the interferences that are produced across different floors, so the interferences will have their origin in other devices from the same floor. Fig. 11 shows the results of our study. We can conclude that the effect of interferences received from other floors is significant enough to deserve consideration. For that reason, the use of planar layouts to model Wi-Fi deployments in multi-floor scenarios like buildings is not recommendable.

5.6. Effect of the propagation model

In this section we evaluate the importance of using different propagation models. In our proposal, we have chosen the indoor transmission loss model defined by ITU-R in the Recommendation P.1238–10 [33]. Obviously, if the setting is a building it is reasonable to use an indoor propagation model. However, as free space propagation models are very widely used, we want to explore whether there are differences in performance when using this type of models instead of an indoor one. As a free space propagation model, we have used the pro-

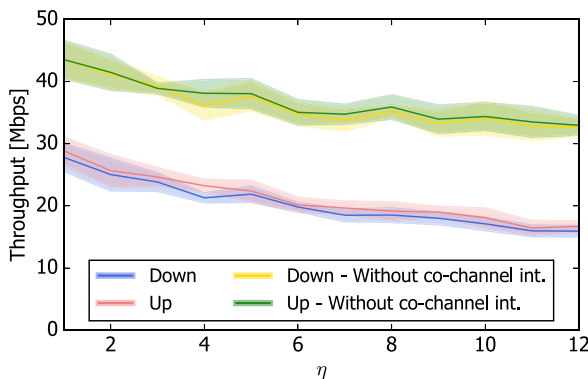


Fig. 10 Effect of the co-channel interference.

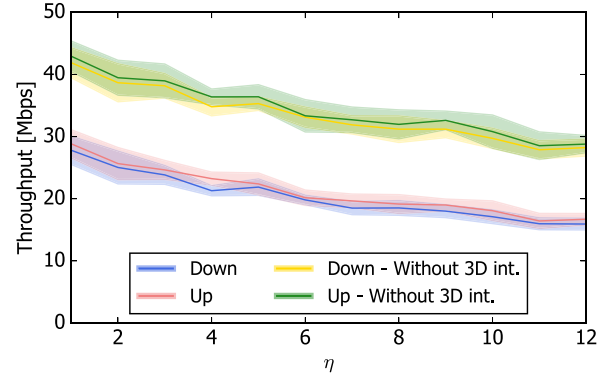


Fig. 11 Effect of using a three-dimensional layout.

posal made in [45], because it considers Wi-Fi in the 2.4 GHz frequency band for free space path loss with line-of-sight (LOS) links with antennas between 1 and 2.5 meters in height, as the height of antennas has an impact in the range where the Fresnel zone contacts the ground plane. The propagation model described in [45] for the 2.4 GHz band is:

$$L_{total} = 7.6 + 40\log_{10}d - 20\log_{10}h_t h_r, \quad (6)$$

being h_t and h_r the height of the transmitting and receiving antennas, respectively. To configure h_t and h_r we have considered them to be 1.5 meters.

Fig. 12 shows the effect of using a free space LOS propagation model instead of the indoor propagation model used in this proposal. We can notice that results significantly differ in both cases. In fact, the throughput achieved when using the free space model propagation is much lower than the throughput obtained in the indoor model. Although losses are higher in the indoor model, it is interesting to note that losses affect not only the desired signal but also the harmful interferences. For that reason, in the free space propagation model the desired signal will be received with higher power, but also the interferences will be higher. Results show that this last behavior prevails over the desired signal, as the final throughput is worse in the free space propagation model.

5.7. Adaptability of the model to other Wi-Fi settings: channel bonding in the 5 GHz frequency band

The experiments shown so far are focused in testing the validity of the proposed model in a well-known and widely studied

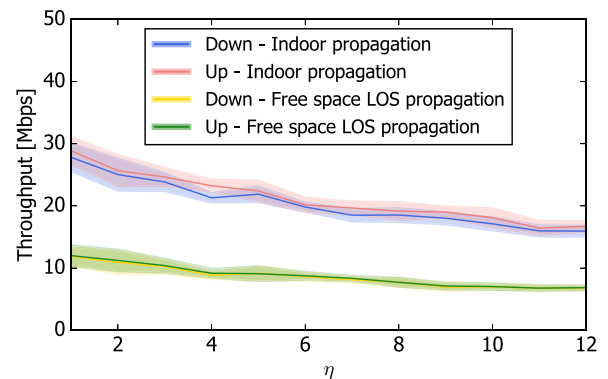


Fig. 12 Effect of using free space LOS propagation model.

Wi-Fi setting (IEEE 802.11n with 20 MHz channels in the 2.4 GHz band). However, one of the key strengths of the model is its adaptability to newer Wi-Fi standards. In this section, we explain how we can easily adapt the proposal to model another paradigmatic example of Wi-Fi setting: the IEEE 802.11ac standard (Wi-Fi 5) using channel bonding in the 5 GHz frequency band. In this setting, the standard allows to use four different bandwidths: 20, 40, 80 and 160 MHz. To adapt the previously used model to this setting we must follow the next steps:

1. *Propagation model.* As the frequency band changes from 2.4 GHz to 5 GHz, the propagation model must be changed accordingly. This change can be easily performed by changing the value of f in Eq. 1. Moreover, as it is stated in the ITU-R Recommendation P.1238–10 [33], losses between adjacent floors are higher in the 5 GHz band. For that reason, in the 5 GHz band we use $L_f(n) = 13n$ dB instead of $L_f(n) = 10n$ dB that we used in the 2.4 GHz frequency band.
2. *Modulation, coding and throughput.* IEEE 802.11ac standard defines a new set of MCS with a certain throughput for each one. As discussed above, we will be able to select a certain MCS level as a function of the experienced SINR. In Table 4 we show the different MCS that can be selected in IEEE 802.11ac when a GI of 800 ns is used. Table 4 also shows the throughput obtained for each MCS and the required SINR thresholds for each MCS, according to [46].
3. *Number of channels and thermal noise.* In addition to the channel width, we also need to specify the number of available channels. Of course, the number of available channels strongly depends on the channel width, i.e. on the way channel bonding is performed. As the channelization for Wi-Fi networks in the 5 GHz frequency varies with the world region, we have chosen the one used in USA. In that channelization the channels are orthogonal and we can use up to 25 channels of 20 MHz width, up to 12 channels of 40 MHz, up to 6 channels of 80 MHz or up to 2 channels of 160 MHz (after adding channel 144 in year 2014 to the original 802.11ac specification). Finally, and depending on the channel width chosen, we must configure thermal noise according to Eq. 5.

With the model adapted to IEEE 802.11ac, as described above, we have evaluated the average throughput obtained by STAs under different types of channel bonding schemes, ranging from not using channel bonding (i.e. with 20 MHz channels) to the use of 160 MHz channels, which are the widest possible channels available in IEEE 802.11ac. Moreover, we focus on the 5 GHz frequency band. Finally, although we have evaluated both upstream and downstream throughputs, we only show results for the downlink in Fig. 13, as the conclusions for both are the same. Results show that the best choice in the evaluated setting for all the values of STA density (η) is to use 80 MHz channels, as it is the best trade-off between bandwidth and number of channels. In fact, 80 MHz channels are probably the most widely used in the 5 GHz frequency band. On the other hand, although 160 MHz channels would allow for higher throughputs, the lower number of available channels increases the interferences as we must reuse them very frequently and in closer places, so the experienced throughput is lower. It is also interesting to highlight the slope of the 40 MHz and, specially, 20 MHz curves, because, under these settings, there are enough orthogonal channels to avoid the use of interfering channels in close APs.

5.8. Effect of external interferences: the Bluetooth case

Another interesting setting where our proposed model can be extended and be easily used is the consideration of external interferences that affect Wi-Fi receptions. This phenomenon occurs as IEEE 802.11 networks operate in unlicensed frequency bands. And this behavior can become especially noticeable in the 2.4 GHz frequency band, where several technologies coexist. Although other interference sources are also possible, a prominent and interesting case is the presence of Bluetooth communications close to Wi-Fi devices that can affect their performance. In this section we have extended our model to include external interferences due to close Bluetooth devices. More concretely, we study how Wi-Fi performance changes when there is a Bluetooth device transmitting at a certain distance from a STA. However, as the diversity of Bluetooth devices in terms of transmission power is very high, we must take into consideration the different power classes of Bluetooth devices, as specified in Table 5 [47]. Moreover, due to the frequency hopping operation of Bluetooth,

Table 4 Relation between MCS, SINR and throughput in Wi-Fi 5 with 800 ns GI.

MCS	Mod. scheme	Coding rate	20 MHz		40 MHz		80 MHz		160 MHz	
			Thr. (Mbit/s)	SINR (dB)	Thr. (Mbit/s)	SINR (dB)	Thr. (Mbit/s)	SINR (dB)	Thr. (Mbit/s)	SINR (dB)
0	BPSK	1/2	6.5	[2,5]	13.5	[5,8]	29.3	[8,11]	58.5	[11,14]
1	QPSK	1/2	13	[5,9]	27	[8,12]	58.5	[11,15]	117	[14,18]
2	QPSK	3/4	19.5	[9,11]	40.5	[12,14]	87.8	[15,17]	175.5	[18,21]
3	16-QAM	1/2	26	[11,15]	54	[14,18]	117	[17,21]	234	[21,24]
4	16-QAM	3/4	39	[15,18]	81	[18,21]	175.5	[21,24]	351	[24,27]
5	64-QAM	2/3	52	[18,20]	108	[21,23]	234	[24,26]	468	[27,29]
6	64-QAM	3/4	58.5	[20,25]	121.5	[23,28]	263.3	[26,31]	526.5	[29,34]
7	64-QAM	5/6	65	[25,29]	135	[28,32]	292.5	[31,35]	585	[34,38]
8	256-QAM	3/4	78	≥ 29	162	[32,34]	351	[35,37]	702	[38,40]
9	256-QAM	5/6	-	-	180	≥ 34	390	≥ 37	780	≥ 40

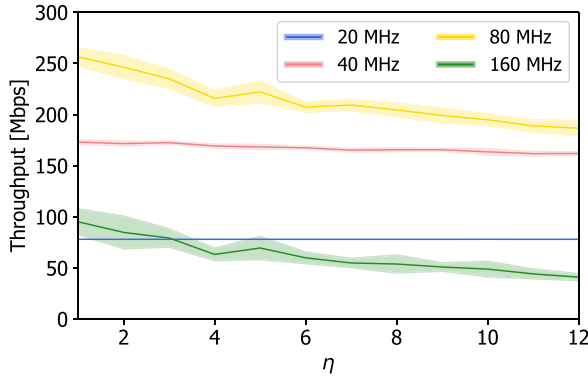


Fig. 13 Effect of channel bonding in IEEE 802.11ac operating in the 5 GHz frequency band.

Bluetooth signals will not interfere IEEE 802.11 signals continuously. Frequency hops occur at a standard hop rate of 1600 hops per second following a pseudo-random sequence, hopping through up to 79 channels (although this quantity can be lower when using Adaptive Frequency Hopping), having each of those channels a bandwidth equal to 1 MHz. To model this pseudo-random behavior of Bluetooth we have averaged the amount of time that a Bluetooth transmission collides with the spectrum used in Wi-Fi. Thus, a Bluetooth communication will collide with a 20 MHz Wi-Fi channel in 20 hops every 79 hops on average, so a Bluetooth communication will interfere with a 20 MHz Wi-Fi channel approximately $20/79 = 25.32\%$ of the time. Moreover, it is also remarkable that a Bluetooth interference does not affect the whole spectrum of a Wi-Fi channel, but only 1 MHz from the total bandwidth of the IEEE 802.11 channel.

Next, we conduct some experiments to show the proposed model capability of considering external interferences into Wi-Fi communications. The main results of this study are shown in Fig. 14, where we show, for a scenario with 6 STAs per AP, how the downlink throughput of STAs is affected when a Bluetooth device is transmitting at a certain distance. As the different transmission power classes are defined in terms of a transmission power interval, we have used the worst case for all of them in terms of interferences. Taking this into account, we have used the highest power for each interval (100 mW for class 1, 2.5 mW for class 2 and 1 mW for class 3). In Fig. 14 we can see that, as expected, when Bluetooth transmitting devices are closer to the Wi-Fi STAs, throughput can be severely affected, especially with the highest power Bluetooth class. And, as Bluetooth interfering devices are further, their effect quickly dispels. Moreover, it is important to highlight that, as expected due to its low transmission power, the effect of class 2 and class 3 Bluetooth devices in Wi-Fi communications can only be noticed at very short distances.

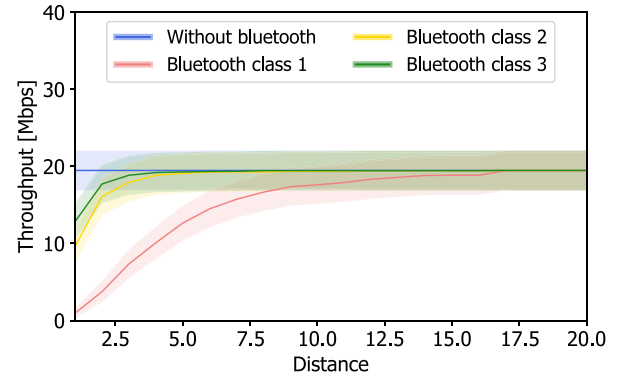


Fig. 14 Effect of Bluetooth interferences in IEEE 802.11.

6. Conclusions and future work

One of the main problems in Wi-Fi network planning and optimization research is the difficulty to compare approaches from different authors, due to the vast diversity of models considered by the different research groups working in this area. The present paper bridges this gap presenting a generic model for signal propagation, goodness computation, and scenario representation in Wi-Fi infrastructure networks, which is flexible and easily reusable by the community. First, we provide a realistic model to represent Wi-Fi settings, including architecture, signal propagation, and interference and throughput computation. Then, we propose a compact graph-based scenario representation and a collection of realistic scenario settings representing residential buildings with several floors (hence three-dimensional scenarios), with varying density of wireless devices (for a total of 120 scenarios). Finally, we have conducted a comprehensive set of validation experiments including a comparison with the well-known discrete event simulator *ns-3*. Results show the consistency of the model and the relevance of some of its design features, like the use of 3-D graphs and the consideration of co-channel interference.

For validation, our proposed model has been particularized for a specific kind of realistic Wi-Fi network deployments. More specifically, the design requirements of the model have been driven by the features of dense and ultra-dense Wi-Fi networks operating in the 2.4 GHz band, as these types of networks are usually the ones that require a more accurate design and optimization to prevent clients from obtaining a very poor quality of service.

For the experimental evaluation of the model, we have made a number of assumptions. We have restricted ourselves to the 2.4 GHz band, indoor residential propagation, and specific co-channel interference values [38]. We have also assumed that clients associate to the access point of their same residential unit. We believe these assumptions to be reasonable and realistic. In any case, the model is flexible enough to support other bands, propagation parameters, and interference assumptions. We believe this model adequately fills the gap for the comparison of different proposals in the earlier and middle stages of research, before fine-tuning with discrete-event simulations.

We believe the two main strengths of our proposal are its realism and adaptability. By taking into account STA density,

Table 5 Power transmission classes defined in Bluetooth [47].

Power class	Power (in mW)	Power (in dBm)
1	(2.5, 100]	(+4, +20]
2	(1, 2.5]	(0, +4]
3	≤ 1	≤ 0

co-channel interference, 3D wireless settings and indoor propagation, the model allows to accurately represent practical realistic scenarios, so that researchers can use it to refine their approaches before carrying out more computation- and time-expensive simulations. Also, as we have shown in the last sections of the paper, the model can be easily adapted to incorporate other technologies or modifications over the existing ones, which allows for an agile exploration of the design space when researching on new methods to improve the efficiency of our exploitation of the wireless spectrum. We envision our model being used for rapid prototyping of new features in wireless controllers, access points and STAs, and even of new protocol extensions and updates [48].

Though the experiments performed with the model and the graphs generated yield satisfactory results, there is still plenty of research to be done in this area. First of all, our model is static, so right now to work with highly changing environments users will need to either abstract that dynamicity (e.g. taking averages) or to work with time “slices” where the properties of the scenario are stationary enough. One of our immediate avenues for future research is to consider dynamic scenarios. In addition, we are interested in exploring different graph generation strategies, not only to account for different building shapes, but also to widen the diversity of the topological properties of the scenario graph, thus enabling studies about how graph metrics impact the relative performance of different coordination approaches, similar to the ones we conducted in [32]. We are also interested in generating the most usual scenarios in the literature and in performing an exhaustive comparison of the most relevant works in those scenarios. Finally, we plan to integrate the model and the analysis code (right now available as supplementary material for this article) in a web service, which will allow to share scenarios and experiment results with the Wi-Fi planning and optimization community more easily, further facilitating in this way that researchers compare and share their advancements.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially funded by Projects PID2019-104855RB-I00/AEI/10.13039/501100011033 and PID2019-104129GB-I00/AEI/10.13039/501100011033, by Project SBPLY/19/180501/000171 of the Junta de Comunidades de Castilla-La Mancha and FEDER, by Projects UCeNet (CM/JIN/2019-031) and WiDAI (CM/JIN/2021-004) of the Comunidad de Madrid and University of Alcalá, and by H2020-MSCA-RISE project 734922 – CONNECT. The publication is also part of project TED2021-131387B-I00 funded by MCIN/AEI/10.13039/501100011033 and by the European Union “NextGenerationEU”/PRTR and of project PID2021-123168NB-I00 funded by MCIN/AEI/10.13039/501100011033 / FEDER, UE.

Appendix A. Supplementary materials

Our aim is for the proposed model to be useful to the research community to evaluate and validate their approaches for Wi-Fi network planning, coordination, and optimization. In a sense, we are trying to provide the model we would gift our “past-selves” seven years ago to avoid a number of headaches and bottlenecks. It is our hope that readers can benefit of this model to facilitate their research progress in this fascinating and promising area.

The graph files and the corresponding EPS figures can be downloaded from <https://doi.org/10.5281/zenodo.5703320>. The code (in Python) we have used to evaluate throughput in a given setting is available upon request. All the material is licensed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>), which means that you can use it and adapt it as long as you cite this paper as its original source, which we would of course greatly appreciate.

References

- [1] H.A. Omar, K. Abboud, N. Cheng, K.R. Malekshan, A.T. Gamage, W. Zhuang, A survey on high efficiency wireless local area networks: Next generation WiFi, *IEEE Commun. Surv. Tutor.* 18 (4) (2016) 2315–2344.
- [2] S. Chiochan, E. Hossain, J. Diamond, Channel assignment schemes for infrastructure-based 802.11 WLANs: A survey, *IEEE Commun. Surv. Tutor.* 12 (1) (2010) pp.
- [3] E. de la Hoz, J.M. Gimenez-Guzman, I. Marsa-Maestre, D. Orden, Automated negotiation for resource assignment in wireless surveillance sensor networks, *Sensors* 15 (11) (2015) 29547–29568.
- [4] H.-J. Chen, C.-P. Chuang, Y.-S. Wang, S.-W. Ting, H.-Y. Tu, C.-C. Teng, Design and implementation of a cluster-based channel assignment in high density 802.11 WLANs, in: *Network Operations and Management Symposium (APNOMS)*, 2016 18th Asia-Pacific, IEEE, 2016, pp. 1–5.
- [5] Y.M. Kwon, K. Choi, M. Kim, M.Y. Chung, Distributed channel selection scheme based on the number of interfering stations in WLAN, *Ad Hoc Netw.* 39 (2016) 45–55.
- [6] H. Kasasbeh, F. Wang, L. Cao, R. Viswanathan, Generous throughput oriented channel assignment for infra-structured WiFi networks, in: *Wireless Communications and Networking Conference (WCNC)*, 2017 IEEE, IEEE, 2017, pp. 1–6.
- [7] K. Chen, *Performance evaluation by simulation and analysis with applications to computer networks*, John Wiley & Sons, 2015.
- [8] A. Krotov, A. Kiryanov, E. Khorov, Rate control with spatial reuse for Wi-Fi 6 dense deployments, *IEEE Access* 8 (2020) 168898–168909.
- [9] S.M. Kala, V. Sathya, W.K. Seah, H. Yamaguchi, T. Higashino, Evaluation of theoretical interference estimation metrics for dense Wi-Fi networks, in: *2021 International Conference on Communication Systems & NETworkS (COMSNETS)*, IEEE, 2021, pp. 351–359.
- [10] B. Dab, I. Fajjari, N. Aitsaadi, Optimized wireless channel allocation in hybrid data center network based on IEEE 802.11 ad, *Comput. Commun.* 160 (2020) 534–546.
- [11] J.M. Gimenez-Guzman, I. Marsa-Maestre, D. Orden, S. Fernandez, M. Tejedor-Romero, On the benefits of channel

- bonding in dense, decentralized Wi-Fi 4 networks, *Wireless Communications and Mobile Computing* 2022 (8497585) (2022) 1–11.
- [12] J. Martinez-Bauset, J.M. Gimenez-Guzman, V. Pla, Optimal admission control in multimedia mobile networks with handover prediction, *IEEE Wirel. Commun.* 15 (5) (2008) 38–44.
 - [13] J.M. Gimenez-Guzman, M.J. Domenech-Benlloch, V. Pla, V. Casares-Giner, J. Martinez-Bauset, Analysis of a cellular network with user redials and automatic handover retrials, in: *International Conference on Next Generation Wired/Wireless Networking*, Springer, 2007, pp. 210–222.
 - [14] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *Proceedings of the national academy of sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.
 - [15] D. Lopez-Pajares, J. Alvarez-Horcajo, E. Rojas, J.A. Carral, G. Ibanez, Iterative discovery of multiple disjoint paths in switched networks with multicast frames, in: *2018 IEEE 43rd Conference on Local Computer Networks (LCN)*, IEEE, 2018, pp. 409–412.
 - [16] A.U. Chaudhry, R.H. Hafez, J.W. Chinneck, On the impact of interference models on channel assignment in multi-radio multi-channel wireless mesh networks, *Ad Hoc Netw.* 27 (2015) 68–80.
 - [17] E. Ahmed, A. Gani, S. Abolfazli, L.J. Yao, S.U. Khan, Channel assignment algorithms in cognitive radio networks: Taxonomy, open issues, and challenges, *IEEE Communications Surveys & Tutorials* 18 (1) (2014) 795–823.
 - [18] E. De La Hoz, I. Marsa-Maestre, J.M. Gimenez-Guzman, D. Orden, and M. Klein, "Multi-Agent nonlinear negotiation for Wi-Fi channel assignment," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 1035–1043, International Foundation for Autonomous Agents and Multiagent Systems, 2017.
 - [19] A. Modarresi, J. Symons, Modeling and graph analysis for enhancing resilience in smart homes, *Procedia Computer Science* 160 (2019) 197–205.
 - [20] H. Mabed, P. Canalda, F. Spies, Geometry modeling in cellular network planning, *Wireless networks* 20 (6) (2014) 1251–1263.
 - [21] Y. Qu, Goodput Modelling and Optimisation of Channel Assignment For Planning IEEE 802.11 Wireless Backhaul Networks, Victoria University of Wellington, 2018, PhD thesis.
 - [22] K.I. Aardal, S.P. Van Hoesel, A.M. Koster, C. Mannino, A. Sassano, Models and solution techniques for frequency assignment problems, *Ann. Oper. Res.* 153 (1) (2007) 79–129.
 - [23] D. Orden, J.M. Gimenez-Guzman, I. Marsa-Maestre, E. de la Hoz, Spectrum graph coloring and applications to Wi-Fi channel assignment, *Symmetry* 10 (3) (2018) 65.
 - [24] A. Mishra, S. Banerjee, W. Arbaugh, Weighted coloring based channel assignment for WLANs, *ACM SIGMOBILE Mobile Computing and Communications Review* 9 (3) (2005) 19–31.
 - [25] A. Mishra, V. Shrivastava, S. Banerjee, W. Arbaugh, Partially overlapped channels not considered harmful, *ACM SIGMETRICS Performance Evaluation Review*, vol. 34, ACM, 2006, pp. 63–74.
 - [26] J.M. Gimenez-Guzman, I. Marsa-Maestre, D. Orden, E. de la Hoz, T. Ito, On the goodness of using orthogonal channels in WLAN IEEE 802.11 in realistic scenarios, *Wireless Communications and Mobile Computing* 2018 (5742712) (2018) 1–11.
 - [27] O. Jeunen, P. Bosch, M. Van Herwegen, K. Van Doorselaer, N. Godman, S. Latré, A machine learning approach for IEEE 802.11 channel allocation, in: *2018 14th International Conference on Network and Service Management (CNSM)*, IEEE, 2018, pp. 28–36.
 - [28] M. Stojanova, T. Begin, A. Busson, Conflict graph-based model for IEEE 802.11 networks: A divide-and-conquer approach, *Performance Evaluation* 130 (2019) 64–85.
 - [29] G. Capdehourat, P. Bermolen, M. Fiori, N. Frevenza, F. Larroca, G. Morales, C. Rattaro, G. Zunino, Large-scale 802.11 wireless networks data analysis based on graph clustering, *Wireless Pers. Commun.* (2021) 1–29.
 - [30] L. Abdelwedoud, A. Busson, I.G. Lassous, Use of a weighted conflict graph in the channel selection operation for Wi-Fi networks, in: *2021 16th Annual Conference on Wireless On-demand Network Systems and Services Conference (WONS)*, IEEE, 2021, pp. 1–4.
 - [31] A. Chadda, M. Stojanova, T. Begin, A. Busson, I.G. Lassous, Assigning channels in WLANs with channel bonding: A fair and robust strategy, *Comput. Netw.* (2021) 108200.
 - [32] I. Marsa-Maestre, E. de la Hoz, J.M. Gimenez-Guzman, D. Orden, M. Klein, Nonlinear negotiation approaches for complex-network optimization: a study inspired by Wi-Fi channel assignment, *Group Decis. Negot.* 28 (1) (2019) 175–196.
 - [33] R. ITU-R, "1238-10," Propagation data and prediction methods for the planning of indoor radiocommunication systems and radio local area networks in the frequency range, vol. 300.
 - [34] A. Mishra, V. Brik, S. Banerjee, A. Srinivasan, and W.A. Arbaugh, "A client-driven approach for channel management in wireless LANs," in *Infocom*, 2006.
 - [35] A. Voicu, L. Lava, L. Simić, and M. Petrova, "The importance of adjacent channel interference: Experimental validation of ns-3 for dense Wi-Fi networks," pp. 43–52, 11 2017.
 - [36] T. Chrysikos, G. Georgopoulos, S. Kotsopoulos, Site-specific validation of ITU indoor path loss model at 2.4 GHz, in: *2009 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks & Workshops*, IEEE, 2009, pp. 1–6.
 - [37] N. Gupta et al., Analytical approach towards available bandwidth estimation in wireless ad hoc networks, *Wireless Netw.* (2020) 1–26.
 - [38] K.R. Chowdhury, I.F. Akyildiz, Cognitive wireless mesh networks with dynamic spectrum access, *IEEE J. Sel. Areas Commun.* 26 (1) (2008) pp.
 - [39] "IEEE standard for information technology– local and metropolitan area networks– specific requirements– part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications amendment 5: Enhancements for higher throughput," *IEEE Std 802.11n-2009 (Amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008, IEEE Std 802.11r-2008, IEEE Std 802.11y-2008, and IEEE Std 802.11w-2009)*, pp. 1–565, 2009.
 - [40] M. Kim, C.-H. Choi, Hidden-node detection in IEEE 802.11n wireless LANs, *IEEE Trans. Veh. Technol.* 62 (6) (2013) 2724–2734.
 - [41] G.F. Riley, T.R. Henderson, *The ns-3 network simulator*, Modeling and tools for network simulation, Springer, 2010, pp. 15–34.
 - [42] C. Camacho-Gómez, I. Marsa-Maestre, J.M. Gimenez-Guzman, S. Salcedo-Sanz, A coral reefs optimization algorithm with substrate layer for robust Wi-Fi channel assignment, *Soft. Comput.* 23 (23) (2019) 12621–12640.
 - [43] D. Orden, I. Marsa-Maestre, J.M. Gimenez-Guzman, E. de la Hoz, A. Álvarez-Suárez, Spectrum graph coloring to improve Wi-Fi channel assignment in a real-world scenario via edge contraction, *Discrete Applied Mathematics* 263 (2019) 234–243.
 - [44] M. Achanta, "Method and apparatus for least congested channel scan for wireless access points," Apr. 6 2006. US Patent App. 10/959,446.
 - [45] D.B. Green, A. Obaidat, An accurate line of sight propagation performance model for ad-hoc 802.11 wireless LAN (WLAN) devices, in *Communications*, 2002. ICC 2002. IEEE International Conference on, vol. 5, IEEE, 2002, pp. 3424–3428.

- [46] Y.-D. Chen, D.-R. Wu, T.-C. Sung, K.-P. Shih, DBS: A dynamic bandwidth selection MAC protocol for channel bonding in ieee 802.11ac WLANs, in: 2018 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2018, pp. 1–6.
- [47] Core Specification Working Group, "Bluetooth core specification v5.3," 2021.
- [48] I. Marsa-Maestre, J.M. Gimenez-Guzman, M. Tejedor-Romero, E. de la Hoz, and P. Murukannaiah, "Democratic wireless channel assignment: Fair resource allocation in Wi-Fi networks," IEEE Internet Computing, (to appear).

Channel Bonding in Wi-Fi 4

” *The whole is greater than the sum of its parts*

— Aristotle

(Ancient Greek philosopher)

As we mentioned in the previous chapter, Wi-Fi network optimization is a popular research line. In particular, one of the most discussed problems is the impact of channel management in the performance of the network. This does not only refer to frequency channel distribution among access points, but also to techniques like channel bonding, which merges two adjacent channels in order to conform a single frequency channel with a larger bandwidth. In consequence, less channels are available, since they are wider.

Previous works stated that channel bonding is counter-productive in 802.11n (Wi-Fi 4), due to the co-channel interference between partially overlapping frequencies. In this article [Gim+22], we study this technique with the aid of our graph-based model. Here, the importance of edge parameters is critical, due to the effects of co-channel interference.

The results of our experiments using the Wi-Fi graph model reveal that channel bonding is only counter-productive in areas where the 2.4GHz band is limited to 11 channels, like North America. In places where this frequency band allows 13 or more channels (almost any other part of the world), there is a consistent benefit on throughput.

Research Article

On the Benefits of Channel Bonding in Dense, Decentralized Wi-Fi 4 Networks

Jose Manuel Gimenez-Guzman ¹, Ivan Marsa-Maestre ², David Orden ³,
Susel Fernandez ² and Marino Tejedor-Romero ³

¹Communications Department, Universitat Politècnica de València, Valencia, Spain

²Computer Engineering Department, Universidad de Alcalá, Alcalá de Henares, Spain

³Department of Physics and Mathematics, Universidad de Alcalá, Alcalá de Henares, Spain

Correspondence should be addressed to Jose Manuel Gimenez-Guzman; jmgimenez@upv.es

Received 7 July 2021; Revised 24 November 2021; Accepted 21 January 2022; Published 10 February 2022

Academic Editor: Ernestina Cianca

Copyright © 2022 Jose Manuel Gimenez-Guzman et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Channel bonding is a technique first defined in the IEEE 802.11n standard to increase the throughput in wireless networks by means of using wider channels. In IEEE 802.11n (nowadays also known as Wi-Fi 4), it is possible to use 40 MHz channels instead of the classical 20 MHz channels. Although using channel bonding can increase the throughput, the classic 802.11 setting only allows for two orthogonal channels in the 2.4 GHz frequency band, which is not enough for proper channel assignment in dense settings. For that reason, it is commonly accepted that channel bonding is not suitable for this frequency band. However, to the best of our knowledge, there is not any accurate study that deals with this issue thoroughly. In this work, we study in depth the effect of channel bonding in Wi-Fi 4 dense, decentralized networks operating in the 2.4 GHz frequency band. We confirm the negative effect of using channel bonding in the 2.4 GHz frequency band with 11 channels which are 20 MHz wide (as in North America), but we also show that when there are 13 or more channels at hand (as in many other parts of the world, including Europe and Japan), the use of channel bonding yields consistent throughput improvements. For that reason, we claim that the common assumption of not considering channel bonding in the 2.4 GHz band should be revised.

1. Introduction and State of the Art

The huge increase of wireless devices competing for the limited wireless bandwidth [1] has attracted the attention of researchers, since it is an increasingly complex problem. Especially in the case of the 2.4 GHz band, where a greater number of devices and protocols coexist, and in dense, decentralized settings such as residential buildings, we find very inefficient bandwidth usage situations [2]. The community is addressing this challenge in a twofold manner. Some researchers focus on new standards and specifications for high-efficiency wireless local area networks (HEWs) [2]. Others, however, focus on improving the centralized or decentralized coordination of devices and networks using the existing standards.

In the latter case, channel assignment techniques aim to optimize the distribution of channels among the transmitting devices, thus decreasing interference and increasing throughput [3–8]. An additional possibility, which exists since the IEEE 802.11n standard (Wi-Fi 4), is the use of *channel bonding*, which consists of using channels that are wider than the standard 20 MHz to achieve higher performance (higher bandwidth would allow for higher transmission rates, thus increasing throughput). A number of channel bonding techniques have been proposed in the literature for the different IEEE 802.11 standards [9–11]. We are especially interested in the standard IEEE 802.11n in the 2.4 GHz band. However, the general consensus is that using channel bonding in the 2.4 GHz band is not beneficial, since the interference due to the use of wider overlapping channels

jeopardizes the theoretical advantage of having higher maximum bit rates [12]. Consequently, most studies assume other bands such as the 5 GHz band [13–16], and the more recent ones generally assume dynamic channel bonding schemes for the 802.11ax standard [11,17].

Nevertheless, in our opinion, the possibilities of static channel bonding in the 802.11n 2.4 GHz have not been properly analyzed in the literature. Since most of the papers mentioning the limitations of channel bonding in this band directly focus on other bands or technologies (i.e., the 2.4 GHz band is not the focus of the paper), they either state these limitations as a matter of fact, not citing any study to back up the claim [18] or cite other papers which, in turn, state such limitations without the backup of an academic study [12,15,16,19]. There are some references to industrial white papers such as [20], but there the North American 11-channel 802.11 spectrum is assumed, although there are many regions in the world (e.g., Europe or Japan) where more channels are available. Furthermore, some studies directly assume the use of orthogonal channels, and therefore they do not consider interferences between adjacent channels [21], or they use only one wireless station (STA) per access point (AP) [10], while it has been shown that both interference between adjacent channels and STA number and precise placement may have a significant effect on throughput [22]. Taking this into account, we believe that the aforementioned consensus about the goodness of channel bonding in the 2.4 GHz band should be revised, as it is based on former studies [23–25] that concluded that channel bonding causes more harmful problems than it solves but these studies do not represent the density of current Wi-Fi networks.

In this paper, we study the effect of channel bonding in dense, decentralized Wi-Fi 4 scenarios, such as a residential building. The paper contributions can be summarized as follows:

- (i) We describe a graph-based scenario model for Wi-Fi 4 dense decentralized networks, using realistic indoor signal propagation and interference models, as well as the precise location and interference between all wireless devices (both access points (APs) and stations (STAs)), in order to compute the throughput. To the best of our knowledge, this is the first time that such an accurate model is used in the context of channel bonding (Section 2).
- (ii) We provide a three-dimensional realistic setting for a decentralized Wi-Fi 4 deployment in a residential building. For this setting, we generate 60 scenarios for different STA densities and placements (Section 2.1).
- (iii) We conduct an in-depth evaluation with the aforementioned model and setting, first for the classic 11-channel Wi-Fi 4 settings (as in, e.g., North America) and then for 13-channel Wi-Fi 4 (as in, e.g., Europe) (Sections 2.3 and 2.4).

Our results show that, on average, the use of channel bonding in an 11-channel Wi-Fi 4 setting yields a decrease in

performance, although there may be some clusters of STAs reaping significant benefits at the expense of the others, which yields fairness concerns. This essentially matches the premises and conclusions in the consensus about channel bonding so far. However, for the 13-channel setting, our results show a consistent advantage of using channel bonding, contrary to the previous belief. The potential fairness issues remain, which opens interesting avenues for future work as we discuss in Section 4.

2. Wi-Fi 4 Network Model

2.1. Wi-Fi Networks. IEEE 802.11 networks, commercially known as Wi-Fi, are the most widespread technology to deploy wireless local area networks (WLANs). Although Wi-Fi networks can operate in ad hoc and infrastructure modes, in this work we focus on the infrastructure mode, as it is the most widely used. In this operating mode, all the communications occur between access points (APs) and their associated stations (STAs), so if two STAs want to communicate to each other, this communication must go through an AP.

One of the main features of Wi-Fi networks is that this type of networks operates in unlicensed frequency bands. Among these frequency bands, we can highlight the spectrum around 2.4 GHz and the spectrum around 5 GHz. Although the 5 GHz band offers higher bandwidth and throughput, the 2.4 GHz is still the most widely used frequency band due to its better coverage and its compatibility with more legacy equipment. To overcome the limitations in bandwidth, Wi-Fi standards have proposed the use of wider frequency channels, which is called *channel bonding*. More specifically, IEEE 802.11n (Wi-Fi 4) proposes the use of 40 MHz channels instead of the classic 20 MHz ones. Later standards open the possibilities of channel bonding to other bands and wider channel widths, up to 160 MHz. In this paper, we will focus on the Wi-Fi 4 standard in the 2.4 GHz band.

2.2. Graph Modeling. To evaluate the effect of channel bonding in dense Wi-Fi networks, we make use of graph models that accurately represent the peculiarities of this type of networks. In fact, our models represent a set of independent Wi-Fi networks spatially distributed and modeled using geometric graphs. A graph can be defined as a set of vertices V and a set of edges E between them, $E \subseteq \{(u, v) | u, v \in V\}$. In our case, we consider geometric graphs, because the spacial positions of both APs and STAs (which will be the two kinds of vertices in our graphs) have a strong influence in the performance of the network [22]. In our graph model, we will also have two types of edges, one type representing the association between STAs and APs and the other type representing the interfering signals between wireless devices of different networks. To model interferences, we use an activity factor to account for the fact that STAs and APs do not transmit continuously, and we assume higher ψ for APs. Although in this paper we consider 3D graphs, for the sake of an easier visualization, Figure 1 shows

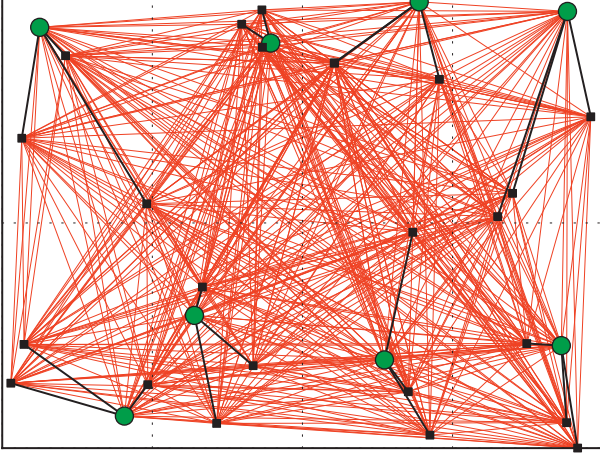


FIGURE 1: Example of Wi-Fi networks modeled with a graph.

a 2D example of the Wi-Fi layout for a single floor of a building, composed of 8 flats, 8 APs (one AP per flat), and 24 STAs (3 STAs per AP). We represent the APs as green circles and the STAs as black squares. Regarding edges, black segments represent the associations between APs and STAs, while red segments represent the interfering signals.

We have made use of a graph-based model for the following reasons. Although we have used discrete event simulators in the context of wireless networks in the past [26], we have noted that those papers that study dense Wi-Fi networks using simulation are difficult to replicate, especially because of the effect of the interferences between adjacent channels [27], which are not negligible at all in dense Wi-Fi settings. For that reason, we chose a graph model that can capture a high number of Wi-Fi network features (not with the precision of simulation models) but is faster and easier to replicate and to use by other researchers for comparison.

2.3. Propagation, Interferences, SINR, and Throughput Computation. Once we have the graph that represents the Wi-Fi layout, we must define how we can compute the achieved throughput for each STA. First of all, it is important to emphasize that the geometry of the problem defines the distances of the different Wi-Fi elements, so with a proper propagation model, we will be able to compute the received signals from the different Wi-Fi elements, being either the desired signal or, mostly, interferences. As we focus on indoor Wi-Fi environments (dense Wi-Fi networks are usually indoor networks), we have used the propagation model defined by the ITU-R in the Recommendation P.1238-10 [28], as it assumes that STAs and APs are in the same building, which will be our testing scenario. Moreover, the ITU-R propagation model also considers losses across different building floors. In [28], propagation losses (in dB) are defined by

$$L_{\text{total}} = 20 \log_{10} f - 28 + N \log_{10} d + L_f(n), \quad (1)$$

with f being the frequency expressed in MHz, N the distance power loss coefficient, d the distance in meters, and $L_f(n)$ the floor penetration factor when signal goes across n

floors. For the 2.4 GHz frequency band, [28] defines $N = 28$ in residential environments, although it is admitted that propagation through walls increases this value considerably. Therefore and according to [29], we have considered $N = 28$ when $d < 16$ meters and $N = 38$ for $d \geq 16$ meters. Finally, according to [28], the losses across two floors when using concrete are 10 dB, so we have considered $L_f(n) = 10n$.

After computing the propagation losses, we can compute the signal power (expressed in dBm) received by an AP or STA i from another Wi-Fi element j as

$$P_r^{j \rightarrow i} = P_t^j + G_j + G_i - L_{\text{total}}, \quad (2)$$

where P_t^j represents the transmission power of j (in dBm) and G_j (or G_i) stands for the transmission (or reception) antenna gain (in dB).

Next, we explain how we compute in the model the interferences received at a device i . In general, a device i will receive interferences from all the transmitting devices in the whole network, excepting from the devices that belong to its same cluster, as their communications are coordinated and do not interfere. Note that a cluster defines the set made by an AP and all its associated STAs. The power of the interfering signal received at device i from device j ($I_r^{j \rightarrow i}$) will be the power of the received signal from j , i.e., $P_r^{j \rightarrow i}$. However, the interference will only be relevant to the device i to the extent that there is an overlap between the spectrum masks (in the frequency domains) as the communications between devices from the same cluster are coordinated and do not interfere. The model accounts for this overlap by means of parameter κ . If both channels are the same, we will consider a total overlap and $\kappa = 1$. On the contrary, if both channels do not collide in the spectrum (orthogonal channels), we will have $\kappa = 0$. Finally, if both channels partially collide, we will consider values of κ ranging from 0 to 1. In addition, we must also consider that, to account for the interference produced from device j to device i , device j is not making use of the spectrum continuously, from a temporal point of view. That behavior is considered by means of the activity factor ψ introduced in Section 2, which can be either ψ_{AP} or ψ_{STA} depending on whether the interfering source is an AP or a STA, respectively. In a sense, factor ψ represents the CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance) behavior of Wi-Fi networks. Some works [30] have modeled it as a Continuous Time Markov Chain (CTMC), concluding that when a STA or AP wants to transmit a packet with probability P_{STA} (or P_{AP} , respectively), it will succeed with probability P_s . For that reason, ψ_{AP} can be computed as $\psi_{AP} = P_{AP} \cdot P_s$, and, equivalently, $\psi_{STA} = P_{STA} \cdot P_s$. As an AP is expected to transmit with a higher probability than STAs, $P_{AP} > P_{STA}$, and therefore $\psi_{AP} > \psi_{STA}$. In summary, we can compute the interference produced by a device j to a device i ($I_r^{j \rightarrow i}$), in a linear scale, by considering the power of the received signal ($P_r^{j \rightarrow i}$), the frequency overlap of their transmission/reception channels (κ), and the activity factor that accounts for the fraction of time during which the interference is being produced (ψ):

$$I^{j \rightarrow i} = P_r^{j \rightarrow i} \cdot \psi \cdot \kappa. \quad (3)$$

From the computation of the desired signal and all the interferences, for a specific STA, it is straightforward to compute the Signal-to-Interference-plus-Noise Ratio (SINR), as it is the quotient of the received power of the signal from its associated AP divided by the sum of the power of all the interfering signals plus the thermal noise. Note that the thermal noise depends on the channel bandwidth, so we consider its value to be -101 dBm for 20 MHz channels and -98 dBm when using channel bonding (with 40 MHz channels).

Finally, to compute the downlink throughput perceived by a STA, we must use the SINR together with the modulation and coding scheme (MCS) used. Depending on the SINR, Wi-Fi 4 [31] defines a specific MCS to be used, which in turn determines the throughput achieved by the STA. As the SINR is higher, it is possible to use modulations with a higher number of bits per symbol and coding schemes with less redundancy. These predefined MCSs, together with the throughput of each MCS for 20 MHz and 40 MHz channels, as defined in the standard [31], are shown in Table 1. Moreover, the table also shows the different SINR thresholds that determine the use of a specific MCS, according to [32].

2.4. Channel Assignment. One of the main configuration challenges in Wi-Fi networks is the choice of the channel to operate in, as defined in Section 1. There have been many works [3,33–35] focused on channel assignment for different Wi-Fi networks. However, channel assignment in uncoordinated Wi-Fi networks is usually based on a local decision based on using the channel where the perceived interference power is minimal [36], so this will be the channel assignment technique considered in this work. More specifically, in the channel selection algorithm, we have considered that each AP periodically scans the spectrum and chooses the channel where it detects the minimum power of interfering signals. This procedure operates asynchronously among the APs changing the order in which the different APs scan the environment. Note that this channel selection procedure represents the usual situation when a user sets up his/her AP leaving the channel selection to a decision of the AP, typically using the option called “Auto” instead of forcing the use of a specific channel. Moreover, as is commonly accepted and as has been suggested in previous works like [22], we restrict the possible channels to be used to the orthogonal channels, so they do not interfere with each other. However, the width of the 2.4 GHz in North America does not allow to use two 40 MHz channels that are totally orthogonal (i.e., nonoverlapping), so we have considered a case where there is some interference between the most separate channels in the spectrum. This will be described in detail in Section 3.3.

3. Performance Evaluation

In this section, we provide an in-depth evaluation of channel bonding in Wi-Fi 4 when operating in the 2.4 GHz frequency band. After a description of the real-world model we have

considered, we perform a validation of the model using the well-known ns-3 simulator [37]. Then, we study channel bonding in two different settings. First, we consider the spectrum that can be used in North America, consisting of 11 channels with 20 MHz width each. For the sake of simplicity, we will name this setting *2.4 GHz USA*. Second, we consider the setting where there are 13 possible channels in the frequency band, as in many parts of the world including Europe. We will name this last setting *2.4 GHz Europe*. Finally, we conduct an analysis of fairness when using channel bonding.

3.1. Experimental Setting. The evaluation of channel bonding has been performed in a three-dimensional realistic setting that represents a five-floor residential building. This scenario is a typical example of a dense uncoordinated Wi-Fi 4 network. The dimensions of the building are $40 \times 30 \times 15$ meters (respectively, length, width, and height; thus, each floor has a height of 3 meters). Each floor has 8 different flats in a 4×2 arrangement. Regarding the distribution of Wi-Fi networks, we consider that each flat has a single AP and a number η of STAs attached to that AP. Note that all the STAs from a flat are attached to the AP from the same flat, which can be the closest AP or not. Moreover, we have considered a wide range of density of STAs in this setting, ranging from $\eta = 1$ STA per AP to $\eta = 12$ STAs per AP. The position of each AP and associated STAs is limited to its flat, with its position in the x - and y -axis being randomly distributed according to a uniform distribution. However, in the z -axis, each AP and each STA is randomly distributed with a normal distribution with a mean of 1.5 meters and a standard deviation of 0.5 meters, bounded to the limits of the floor. To sum up, all the scenarios under study consist of $8 \times 5 = 40$ flats and their corresponding 40 APs and a number of STAs ranging from 40 (when $\eta = 1$) to $12 \times 40 = 480$ (when $\eta = 12$). Finally, for each specific layout, we have considered 5 different settings to account for the randomness in the deployment of the different Wi-Fi elements, for a total of 60 scenarios. Figure 2 shows a graphical representation of two of the scenarios under study, where, for the sake of clarity, we only show the association between APs and STAs.

Finally, Table 2 defines the values used for the main parameters needed to compute the throughput, which in all cases are typical or reasonable values.

3.2. Model Validation. For validation purposes, in this section we include a comparative evaluation of the results obtained using our proposed model with respect to the equivalent results obtained using a discrete event simulator. More specifically, we have chosen the well-known ns-3 simulator [37]. The reference setting for this validation consists of a single AP and a single STA (attached to that AP) positioned at different distances. As our model computes the highest reachable throughput that a STA is able to obtain, in the simulator we have considered a greedy traffic source that emits UDP datagrams with a rate higher than the maximum throughput that the technology permits. To make the results

TABLE 1: Relation between MCS, SINR, and throughput in Wi-Fi 4 with mandatory 800 ns guard interval (GI) [31].

MCS index	Modulation scheme	Coding rate	Throughput for 20 MHz (Mbit/s)	Throughput for 40 MHz (Mbit/s)	SINR range (dB) [32]
0	BPSK	1/2	6.5	13.5	[6.8, 7.9)
1	QPSK	1/2	13.0	27.0	[7.9, 10.6)
2	QPSK	3/4	19.5	40.5	[10.6, 13.0)
3	16-QAM	1/2	26.0	54.0	[13.0, 17.0)
4	16-QAM	3/4	39.0	81.0	[17.0, 21.8)
5	64-QAM	2/3	52.0	108.0	[21.8, 24.7)
6	64-QAM	3/4	58.5	121.5	[24.7, 28.1)
7	64-QAM	5/6	65.0	135.0	> 28.1

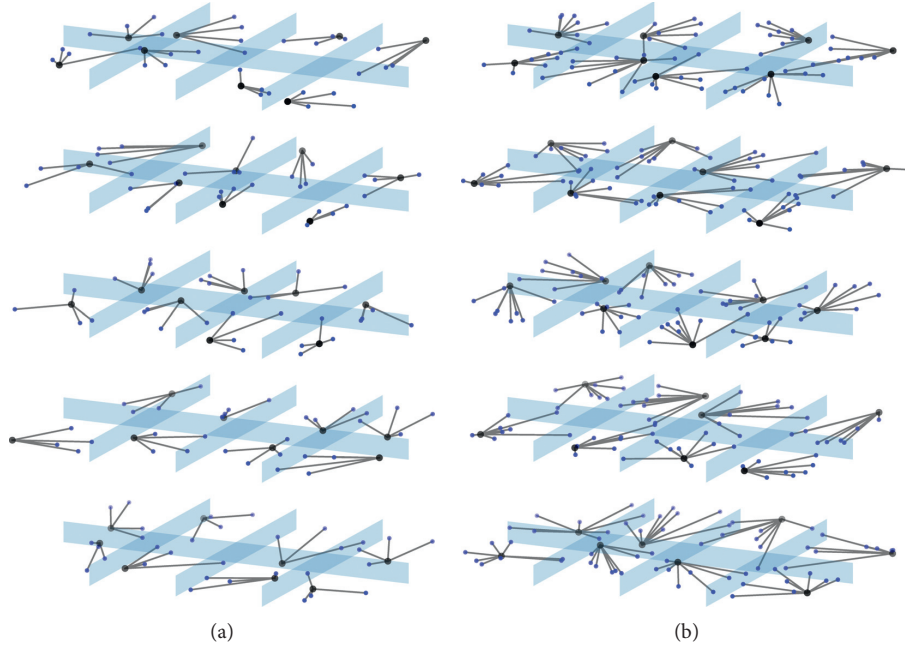
FIGURE 2: Examples of scenarios. (a) $\eta = 4$. (b) $\eta = 8$.

TABLE 2: Summary of parameters.

Parameter	Value
P_t	30 mW
G_t	0 dB
G_r	0 dB
Ψ_{AP}	0.5
Ψ_{STA}	0.1

comparable, we have used in ns-3 the same indoor propagation model, i.e., ITU-R P.1238-10, and we have configured the Wi-Fi manager in the simulator according to the settings used in our experiments.

The validation of our model has been conducted in a two-step procedure. First, at the physical level, we compare the received SINR, as is shown in Figure 3. We can see that both curves totally coincide, validating that our model and the ns-3 simulator obtain the same SINR. As a second step, we study the throughput obtained by the STA with our model and the simulator, as is shown in Figure 4. In that figure, we notice that the shapes of the curves coincide in both cases. However, there is a clear offset between both

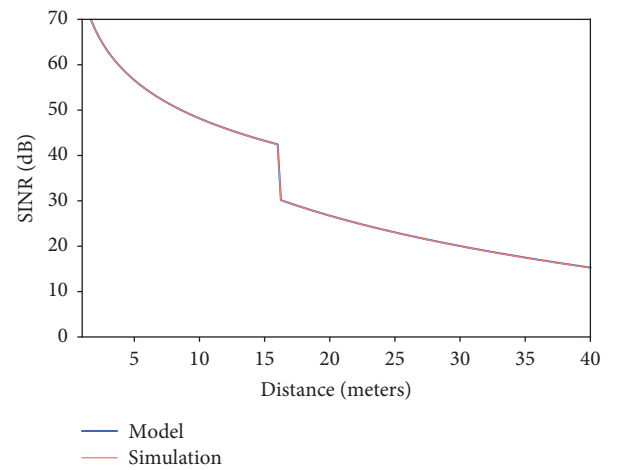


FIGURE 3: SINR obtained by our proposed model and by the ns-3 simulator.

curves. This behavior is due to the fact that our model measures the physical throughput, while ns-3 computes the throughput at the application layer (usually called goodput),

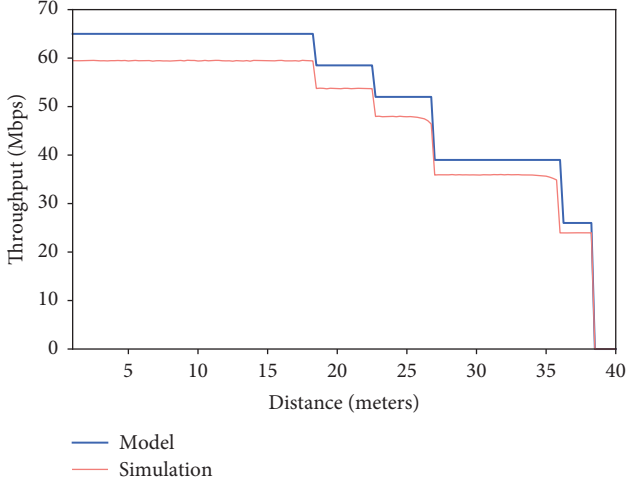


FIGURE 4: Throughput obtained by our proposed model and by the ns-3 simulator. *Note.* The model gives throughput at the physical layer, while the simulator measures goodput (application layer throughput).

with the difference between both values being the overhead introduced by the different layers of the protocol stack. That is the reason of measuring a higher throughput in our model, although both throughput measures coincide.

3.3. Channel Bonding in the 2.4 GHz USA Frequency Band.

In this section, we evaluate the effect that channel bonding has on the throughput perceived by users when they operate in the 2.4 GHz frequency band with 11 nonorthogonal channels, which is the situation occurring in North America. In this setting, when using 20 MHz channels, there are three different orthogonal channels, being channels 1, 6, and 11. However, since the spectrum band goes from 2401 MHz to 2473 MHz, we cannot use two different 40 MHz orthogonal channels. Therefore, by placing one 40 MHz channel in the lowest part of the spectrum and another 40 MHz channel in the highest part, both channels will collide in the frequency band between 2433 and 2441 MHz. For that reason, we have considered that the interference index (κ) when using two 40 MHz channels in the 2.4 GHz USA frequency band is $(2441 - 2433)/40 = 0.2$. In other words, both channels, as they cannot be orthogonal, collide with a factor of $\kappa = 0.2$, producing interferences to each other. Finally, for the sake of completeness, we have also considered the situation where we only use one 40 MHz channel in the 2.4 GHz USA frequency band, since this is the only possibility for totally orthogonal channels in this setting. Figure 5 shows the average downlink throughput and 95% confidence intervals that users can achieve when we consider either three orthogonal 20 MHz channels, two (nonorthogonal) 40 MHz channels, or one orthogonal 40 MHz channel. For each value of η , the average throughput has been computed averaging the five different deployments that we have considered for each value of η and running 10 executions for each of those five deployments. The rationale for performing 10 runs for each setting is that channel assignment technique used (as described in Section 3.4) is not deterministic (except when we make use of a unique 40 MHz

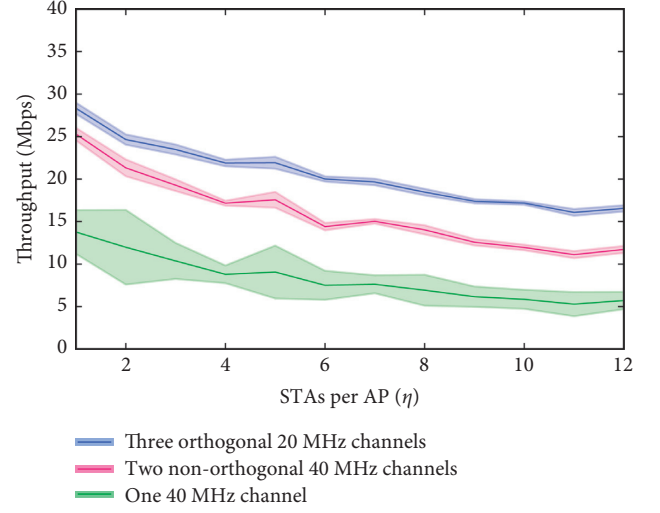


FIGURE 5: Comparison of average throughput in the 2.4 GHz USA frequency band.

channel). As it could be expected, the achieved throughput decreases as the density of STAs (η) increases. Moreover, results show that, on average, in the 2.4 GHz USA frequency band, it is not recommended that channel bonding is used, as the gain that can be obtained by using channels with a higher bandwidth has a lower effect than that of having a fewer number of orthogonal channels, even when the density of users is low ($\eta = 1$). Additionally, we can also conclude that it would be better to use two 40 MHz channels (although they are not completely orthogonal) than to use only one 40 MHz channel.

Now, we perform a more in-depth analysis of the throughput that STAs can achieve individually. In Table 3 we show, for the different densities of STAs (η) under consideration, the percentage of STAs that increase (+), keep the same (=), or decrease (−) their downlink throughput when using channel bonding with two nonorthogonal 40 MHz channels, compared to the situation where three orthogonal 20 MHz channels are used. Moreover, the table also shows the average increase (or decrease) in throughput for the STAs that get better (or worse) performance when using channel bonding with two 40 MHz channels. Inspecting Table 3, we observe that the percentage of STAs that improve their throughput is lower than the percentage of STAs that get a worse throughput, even with the lowest density of STAs, where only 26.5% of STAs fare better in terms of bandwidth and 65% fare worse. This confirms the common belief that, in dense scenarios, it is not recommended that channel bonding is used in the 2.4 GHz USA frequency band. In addition, Table 3 shows that, even when the majority of STAs get a worse throughput, the average gain for the “improving” STAs is higher than the loss for the “losing” STAs, which can create fairness issues and misalignment of incentives in network management for these settings. Since the use of channel bonding is local to APs, the managers of some of the networks could unilaterally decide to transmit in 40 MHz channels at the expense of the networks nearby. Even for those networks which, on average, “lose” throughput by using channel bonding, the fact that when the

TABLE 3: Percentage of STAs that increase (+)/keep (=)/decrease (–) their throughput (and average increase/decrease) using channel bonding with two nonorthogonal 40 MHz channels in the 2.4 GHz USA setting.

η	% STA			Increase/decrease (Mbps)	
	+	=	–	Increase	Decrease
1	26.50	8.50	65.00	20.29	–13.27
2	23.75	7.75	68.50	20.42	–12.49
3	21.83	9.50	68.67	19.00	–12.36
4	18.38	9.25	72.38	19.79	–11.84
5	17.90	11.10	71.00	20.02	–11.56
6	15.83	13.08	71.08	17.59	–11.90
7	16.29	14.36	69.36	17.70	–11.08
8	14.75	13.88	71.38	17.51	–10.45
9	13.56	15.83	70.61	17.31	–10.89
10	12.95	17.40	69.65	17.66	–10.89
11	12.64	18.45	68.91	15.95	–11.05
12	12.67	18.88	68.46	17.15	–10.48

load in the neighboring networks is low, the effective throughput is higher (because of using more bandwidth) may make managers choose to use channel bonding, thus hampering average performance for the network.

We now perform a similar analysis for the comparison of the gain of using one 40 MHz channel with respect to using three 20 MHz orthogonal channels (Table 4). As expected, the use of only one 40 MHz channel is not recommendable, and the number of users that can improve their throughput is very low, ranging from 8% ($\eta = 1$) to 2.73% ($\eta = 10$). We see again that the average throughput increase for “winning” STAs is higher than the decrease for “losing” STAs, but this difference is not as remarkable as in the case of two non-orthogonal 40 MHz channels.

Finally, Figures 6 and 7 show a ridge plot to evaluate the difference in the throughput that each STA can obtain when using two nonorthogonal 40 MHz channels instead of three orthogonal 20 MHz channels (Figure 6) or one 40 MHz channel instead of three orthogonal 20 MHz channels (Figure 7). Note that both figures represent the probability density functions expressed as a Kernel Density Estimation (KDE). In both figures, we observe that the number of STAs that decrease their throughput when using channel bonding (the area to the left of the vertical line at 0) is higher than the number of STAs that improve their throughput (the area to the right). Moreover, the tail on the right side of each KDE is longer than the tail on the left side, reflecting that there are STAs which greatly increase their throughput with channel bonding. Finally, inspecting the figures as the density of the Wi-Fi scenarios increases (i.e., moving vertically from bottom to top), we conclude that channel bonding in the 2.4 GHz USA frequency band is even a worse choice when the density of the Wi-Fi network increases.

3.4. Channel Bonding in the 2.4 GHz Europe Frequency Band.

In this section, we perform a similar evaluation to the one provided in the previous section, but now focusing on the 2.4 GHz frequency band where there are at least thirteen

TABLE 4: Percentage of STAs that increase (+)/keep (=)/decrease (–) their throughput (and average increase/decrease) using channel bonding with one 40 MHz channel in the 2.4 GHz USA setting.

η	% STA			Increase/decrease (Mbps)	
	+	=	–	Increase	Decrease
1	8.00	8.50	83.50	23.38	–19.57
2	7.25	7.75	85.00	24.07	–17.13
3	6.17	9.50	84.33	22.51	–17.27
4	5.13	9.25	85.63	21.63	–17.11
5	4.70	11.10	84.20	22.60	–16.89
6	4.33	13.00	82.67	17.40	–16.23
7	4.00	14.21	81.79	20.47	–16.06
8	3.31	13.81	82.88	19.12	–15.15
9	3.17	15.78	81.06	19.46	–15.26
10	2.70	17.30	80.00	20.46	–15.07
11	2.73	18.36	78.91	14.88	–14.98
12	2.83	18.88	78.29	16.74	–14.81

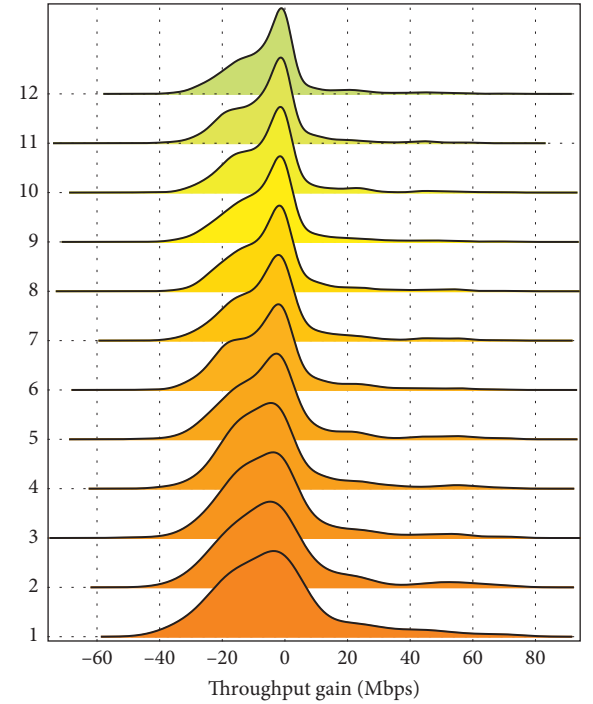


FIGURE 6: Density of the throughput gain of using channel bonding with two 40 MHz channels in the 2.4 GHz USA setting.

20 MHz channels available, as in most of the world and in particular in Europe. In this case, it is possible to use two 40 MHz orthogonal channels, so we will compare two situations: using two orthogonal 40 MHz channels and using three orthogonal 20 MHz channels. Figure 8 shows the average downstream throughput achieved by STAs in both situations. As opposed to the behavior of the 2.4 GHz USA frequency band, in this 2.4 GHz Europe frequency band, the average throughput achieved increases when using channel bonding. Although the advantage of using channel bonding diminishes with the density of STAs (η), it is always advantageous even in the more dense scenarios.

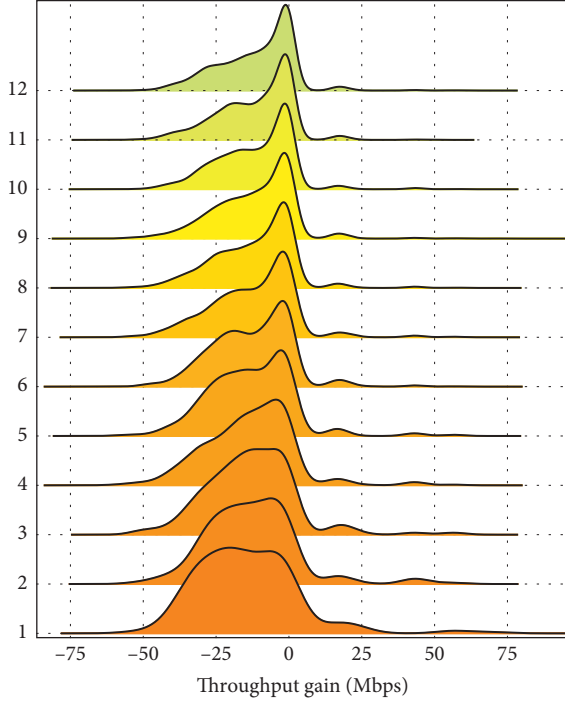


FIGURE 7: Density of the throughput gain of using channel bonding with one 40 MHz channel in the 2.4 GHz USA setting.

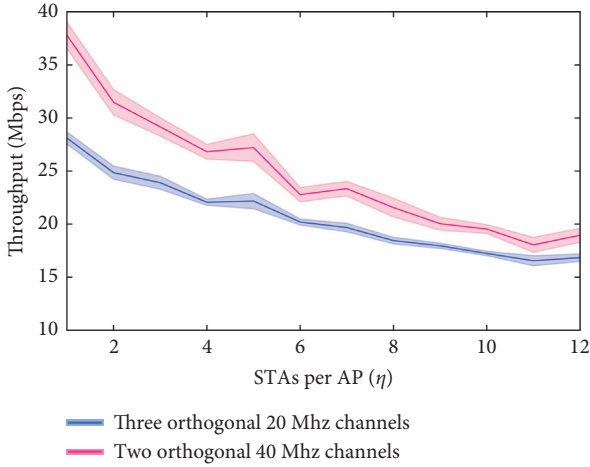


FIGURE 8: Comparison of average throughput in the 2.4 GHz Europe frequency band.

Table 5 shows the number of STAs that are able to increase their throughput when using channel bonding. In general, for the lowest values of η , the percentage of STAs that increase their throughput is higher than the percentage of STAs that decrease it, but this is not true when the density of STAs increases. Furthermore, the average throughput gain of those STAs that improve their throughput is remarkably higher (around 19 to 23 Mbps) than the decrease of those STAs that get a worse throughput (around 6.5 to 9 Mbps).

Finally, in Figure 9, we show the KDE of the difference in the throughput perceived by STAs when using channel bonding, where we can remark the longer tails on the right of

TABLE 5: Percentage of STAs that increase (+)/keep (=)/decrease (−) their throughput (and average increase/decrease) with two orthogonal 40 MHz channels in the 2.4 GHz Europe setting.

η	% STA			Increase/decrease (Mbps)	
	+	=	−	Increase	Decrease
1	55.00	8.50	36.50	23.52	−9.03
2	45.75	7.75	46.50	22.42	−8.20
3	44.33	9.33	46.33	20.48	−7.33
4	41.38	9.25	49.38	20.25	−7.24
5	41.30	11.10	47.60	19.77	−6.72
6	34.33	13.08	52.58	20.68	−7.05
7	35.36	14.21	50.43	19.93	−6.59
8	32.31	13.94	53.75	20.31	−6.10
9	32.11	15.72	52.17	18.99	−6.56
10	28.50	17.40	54.10	19.50	−6.38
11	31.14	18.23	50.64	18.39	−6.66
12	28.96	18.88	52.17	19.03	−6.48

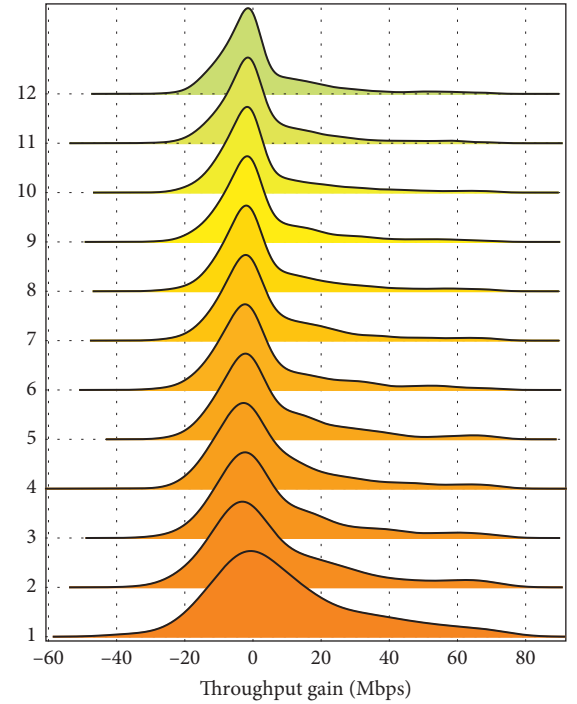


FIGURE 9: Density of the throughput gain of using channel bonding with two orthogonal 40 MHz channels in the 2.4 GHz Europe setting.

the different curves, which shows that the gain of using channel bonding for the “winning” STAs is higher than the loss in throughput for the “losing” ones. Taking this into account, we can conclude that, even when there are a nonnegligible number of STAs that decrease their throughput when using channel bonding, the high increase in an important fraction of STAs makes channel bonding in the 2.4 GHz Europe frequency band worth using.

3.5. Study of Fairness. After the study of the throughput and the incentives that STAs can obtain by using channel

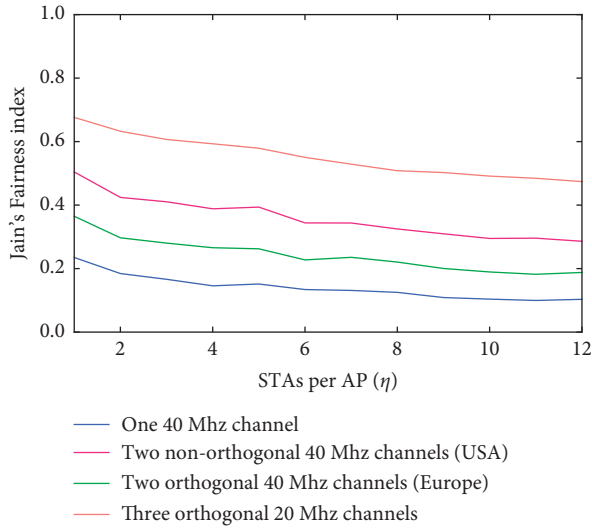


FIGURE 10: Study of fairness.

bonding in different regions, we complete the analysis of channel bonding by studying the fairness of the perceived throughput of the different STAs. The focus is to determine if channel bonding has an effect on fairness. The performance parameter used to measure fairness is the well-known Jain's fairness index [38], defined by

$$f(x) = \frac{[\sum_{i=1}^n Th_i]^2}{\sum_{i=1}^n Th_i^2}, \quad Th_i \geq 0, \quad (4)$$

with Th_i being the downlink throughput perceived by STA i . Jain's fairness index is able to measure the "quality" of the service experienced by the different STAs. If all STAs obtain the same throughput, the fairness index is equal to 1. As the disparity increases, the fairness index goes down to 0, when the system clearly favors selected few users over the rest.

In Figure 10, we show the fairness for the different settings under study. Note that each value of a curve is the average value of the 5 different settings and 10 different executions. Results show that the best fairness is obtained when using three orthogonal 20 MHz channels. This result is due to the fact that the range in throughput is higher when using channel bonding than when not using it. In other words, STAs using channel bonding can obtain a throughput from 0 to 135 Mbps, while the upper value decreases to 65 Mbps when channel bonding is not in use. For this reason, the disparities between STAs when using channel bonding can be higher. The fairness obtained when using two 40 MHz channels is higher in the USA setting than in the Europe setting. Therefore, we can conclude that the advantage in throughput that can be obtained when using two 40 MHz channels in the 2.4 GHz Europe frequency band is at the expense of increasing the disparity between the throughputs obtained by the different STAs.

4. Conclusions

Channel bonding is a technique proposed in Wi-Fi networks since the standard IEEE 802.11n (Wi-Fi 4) to use wider frequency channels to be able to obtain higher throughputs.

However, its use is usually discouraged in the 2.4 GHz frequency band, since it only allows for two 40 MHz orthogonal (or almost orthogonal) channels. However, we found a number of limitations in previous studies on the matter, so we revised that belief for dense, uncoordinated Wi-Fi 4 environments. Our study confirms the previous consensus that it is not advisable to use channel bonding in the 2.4 GHz frequency band with 11 channels which are 20 MHz wide (as in North America). However, contrary to the usual assumption, we show that the use of channel bonding with 40 MHz channels in the 2.4 GHz frequency band with 13 or more 20 MHz channels (the one used in many parts of the world, including Europe) results in an improvement of the average throughput achieved by STAs. Moreover, we show that, even when the number of STAs that decrease their throughput is not negligible, the improvement in throughput experienced by the "winning" STAs is much higher than the decrease in throughput experienced by the "losing" STAs. Hence, channel bonding is worth using not only from the perspective of getting a higher sum of throughputs for the network, but also from the perspective of STAs.

As the decision of using channel bonding lies in the AP but the benefits are for STAs, as a future work, we plan to shift to the STAs (therefore to the users) the decision of whether to use channel bonding or not, since these are ultimately the ones impacted by such decisions. Such a possibility will let us reach more democratic, client-centric configurations, with which we intend to address the fairness issues usually related to channel bonding. Finally, we want to explore the effects of dynamic channel bonding in the 2.4 GHz band, since it could significantly increase the advantage of using channel bonding techniques in Wi-Fi 4.

Data Availability

The graphs used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Jose Manuel Gimenez-Guzman, Ivan Marsa-Maestre, David Orden, and Susel Fernandez were partially funded by Project SBPLY/19/180 501/000 171 of the Junta de Comunidades de Castilla-La Mancha and FEDER and by Project UCeNet (CM/JIN/2019-031) of the Comunidad de Madrid and University of Alcalá. Jose Manuel Gimenez-Guzman, Ivan Marsa-Maestre, and Susel Fernandez were also funded by Project PID2019-104855RB-I00/AEI/10.130 39/501 100 011 033 of the Spanish Ministry of Science and Innovation. David Orden was also partially supported by Project PID2019-104129GB-I00/AEI/10.130 39/501 100 011 033 of the Spanish Ministry of Science and Innovation and by H2020-MSCA-RISE Project 734 922 – CONNECT. Marino Tejedor-Romero was funded by a predoctoral contract from University of Alcalá.

References

- [1] V. Sathya, M. I. Rochman, and M. Ghosh, "Measurement-based coexistence studies of LAA & wi-fi deployments in chicago," *IEEE Wireless Communications*, vol. 28, no. 1, pp. 136–143, 2021.
- [2] H. A. Omar, K. Abboud, N. Cheng, K. R. Malekshan, A. T. Gamage, and W. Zhuang, "A survey on high efficiency wireless local area networks: next generation wifi," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2315–2344, 2016.
- [3] S. Chiochan, E. Hossain, and J. Diamond, "Channel assignment schemes for infrastructure-based 802.11 WLANs: a survey," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 1, 2010.
- [4] E. de la Hoz, J. Gimenez-Guzman, I. Marsa-Maestre, and D. Orden, "Automated negotiation for resource assignment in wireless surveillance sensor networks," *Sensors*, vol. 15, no. 11, pp. 29547–29568, 2015.
- [5] H.-J. Chen, C.-P. Chuang, Y.-S. Wang, S.-W. Ting, H.-Y. Tu, and C.-C. Teng, "Design and implementation of a cluster-based channel assignment in high density 802.11 WLANs," in *Proceedings of the Network Operations and Management Symposium (APNOMS), 2016 18th Asia-Pacific*, pp. 1–5, IEEE, Kanazawa, Japan, October 2016.
- [6] Y. M. Kwon, K. Choi, M. Kim, and M. Y. Chung, "Distributed channel selection scheme based on the number of interfering stations in WLAN," *Ad Hoc Networks*, vol. 39, pp. 45–55, 2016.
- [7] H. Kasasbeh, F. Wang, L. Cao, and R. Viswanathan, "Generous throughput oriented channel assignment for infrastructure WiFi networks," in *Proceedings of the Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, IEEE, San Francisco, CA, USA, March 2017.
- [8] D. Orden, I. Marsa-Maestre, J. M. Gimenez-Guzman, E. de la Hoz, and A. Álvarez-Suárez, "Spectrum graph coloring to improve Wi-Fi channel assignment in a real-world scenario via edge contraction," *Discrete Applied Mathematics*, vol. 263, pp. 234–243, 2019.
- [9] S. H. R. Bukhari, M. H. Rehmani, and S. Siraj, "A survey of channel bonding for wireless networks and guidelines of channel bonding for futuristic cognitive radio sensor networks," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 924–948, 2015.
- [10] S. Barrachina-Muñoz, F. Wilhelmi, and B. Bellalta, "To overlap or not to overlap: enabling channel bonding in high-density wlangs," *Computer Networks*, vol. 152, pp. 40–53, 2019.
- [11] L. Lanante and S. Roy, "Analysis and optimization of channel bonding in dense ieee 802.11 wlangs," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 2150–2160, 2021.
- [12] L. Deek, E. Garcia-Villegas, E. Belding, S.-J. Lee, and K. Almeroth, "The impact of channel bonding on 802.11 N network management," in *Proceedings of the Seventh Conference on Emerging Networking EXperiments and Technologies*, pp. 1–12, Tokyo, Japan, December 2011.
- [13] S.-H. Lim, Y.-B. Ko, C. Kim, and N. H. Vaidya, "Design and implementation of multicasting for multi-channel multi-interface wireless mesh networks," *Wireless Networks*, vol. 17, no. 4, pp. 955–972, 2011.
- [14] L. Deek, E. Garcia-Villegas, E. Belding, S.-J. Lee, and K. Almeroth, "Joint rate and channel width adaptation for 802.11 mimo wireless networks," in *Proceedings of the 2013 IEEE International Conference on Sensing, Communications and Networking (SECON)*, pp. 167–175, IEEE, New Orleans, LA, USA, October 2013.
- [15] J. Fang and I. T. Lu, "Efficient channel access scheme for multiuser parallel transmission under channel bonding in IEEE 802.11ac," *IET Communications*, vol. 9, no. 13, pp. 1591–1597, 2015.
- [16] Y. Daldoul, D.-E. Meddour, and A. Ksentini, "IEEE 802.11 ac: effect of channel bonding on spectrum utilization in dense environments," in *Proceedings of the 2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, Paris, France, May 2017.
- [17] S. Barrachina-Muñoz, F. Wilhelmi, and B. Bellalta, "Dynamic channel bonding in spatially distributed high-density wlangs," *IEEE Transactions on Mobile Computing*, vol. 19, no. 4, pp. 821–835, 2019.
- [18] Z. Khan, H. Ahmadi, E. Hossain, M. Coupechoux, L. A. Dasilva, and J. J. Lehtomäki, "Carrier aggregation/channel bonding in next generation cellular networks: methods and challenges," *IEEE Network*, vol. 28, no. 6, pp. 34–40, 2014.
- [19] D. A. Marendra, G. M. Suranegara, S. Qamar, R. Hakimi, and E. Mulyana, "Emulating software-defined wireless network: bicastrng scenario," in *Proceedings of the 2017 3rd International Conference on Wireless and Telematics (ICWT)*, pp. 76–80, Palembang, Indonesia, July 2017.
- [20] CISCO, *802.11ac: The Fifth Generation of Wi-Fi, Cisco System Technical White Paper*, CISCO, San Jose, CA, USA, 2018.
- [21] L. Xu, K. Yamamoto, and S. Yoshida, "Performance comparison between channel-bonding and multi-channel csma," in *Proceedings of the 2007 IEEE Wireless Communications and Networking Conference*, pp. 406–410, Hong Kong, China, March 2007.
- [22] J. M. Gimenez-Guzman, I. Marsa-Maestre, D. Orden, E. de la Hoz, and T. Ito, "On the goodness of using orthogonal channels in WLAN IEEE 802.11 in Realistic Scenarios," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 5742712, 11 pages, 2018.
- [23] Texas Instruments, *Wlan Channel Bonding: Causing Greater Problems than it Solves*, Texas Instruments, Dallas, Texas, USA, 2003.
- [24] R. Chandra, R. Mahajan, T. Moscibroda, R. Raghavendra, and P. Bahl, "A case for adapting channel width in wireless networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 4, pp. 135–146, 2008.
- [25] V. Shrivastava, S. Rayanchu, J. Yoonj, and S. Banerjee, "11 n under the microscope," in *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, pp. 105–110, Greece, 2008.
- [26] J. Martinez-Bauset, J. Gimenez-Guzman, and V. Pla, "Optimal admission control in multimedia mobile networks with handover prediction," *IEEE Wireless Communications*, vol. 15, no. 5, pp. 38–44, 2008.
- [27] A. M. Voicu, L. Lava, L. Simić, and M. Petrova, "The importance of adjacent channel interference: experimental validation of ns-3 for dense wi-fi networks," in *Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems*, pp. 43–52, Miami, FL, USA, November 2017.
- [28] Rec ITU-R P 1238-8, *Propagation Data and Prediction Methods for the Planning of Indoor Radiocommunication Systems and Radio Local Area Networks in the Frequency Range 300 MHz to 100 GHz*, International Telecommunication Union, Geneva, Switzerland, 2015.
- [29] T. Chrysikos, G. Georgopoulos, and S. Kotsopoulos, "Site-specific validation of itu indoor path loss model at 2.4 GHz," in *Proceedings of the 2009 IEEE International Symposium on*

- World of Wireless, Mobile and Multimedia Networks & Workshops*, pp. 1–6, IEEE, Kos, Greece, 2009.
- [30] Mukta and N. Gupta, “Analytical approach towards available bandwidth estimation in wireless Ad Hoc networks,” *Wireless Networks*, vol. 26, no. 10, pp. 1–26, 2020.
 - [31] IEEE Computer Society, “IEEE standard for information technology– local and metropolitan area networks– Specific requirements– Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications amendment 5: enhancements for higher throughput,” in *Proceedings of the IEEE Std 802.11n-2009 (Amendment To IEEE Std 802.11-2007 as Amended by IEEE Std 802.11k-2008, IEEE Std 802.11r-2008, IEEE Std 802.11y-2008, and IEEE Std 802.11w-2009)*, pp. 1–565, 2009.
 - [32] M. Kim and C.-H. Choi, “Hidden-node detection in ieee 802.11n wireless LANs,” *IEEE Transactions on Vehicular Technology*, vol. 62, no. 6, pp. 2724–2734, 2013.
 - [33] E. De La Hoz, I. Marsa-Maestre, J. M. Gimenez-Guzman, D. Orden, and M. Klein, “Multi-agent nonlinear negotiation for Wi-Fi channel assignment,” in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 1035–1043, Sao Paulo, Brazil, 2017.
 - [34] C. Camacho-Gómez, I. Marsa-Maestre, J. M. Gimenez-Guzman, and S. Salcedo-Sanz, “A coral reefs optimization algorithm with substrate layer for robust wi-fi channel assignment,” *Soft Computing*, vol. 23, no. 23, pp. 12621–12640, 2019.
 - [35] I. Marsa-Maestre, E. de la Hoz, J. M. Gimenez-Guzman, D. Orden, and M. Klein, “Nonlinear negotiation approaches for complex-network optimization: a study inspired by wi-fi channel assignment,” *Group Decision and Negotiation*, vol. 28, no. 1, pp. 175–196, 2019.
 - [36] M. Achanta, “Method and apparatus for least congested channel scan for wireless access points,” US Patent App, 2006.
 - [37] G. F. Riley and T. R. Henderson, “The Ns-3 Network Simulator,” in *Modeling and Tools for Network Simulation*, pp. 15–34, Springer, Berlin, Germany, 2010.
 - [38] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, “A quantitative measure of fairness and discrimination,” Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA, USA, 1984.

Geometric projective mapping for opinion analysis

” Did someone say cookies?

— **Cookie Monster**
(Sesame Street and many websites)

Rapid opinion analysis methods like projective mapping are a useful contribution to the field of opinion analysis, but they have traditionally used statistical techniques. This is an issue for the analysis of large datasets, due to the standard technique used, Multiple Factor Analysis, having cubic computational complexity on the number of participants. This is an issue when the current technologies allow us to gather larger amounts of data in less time with ease. This is why the geometric alternative *SensoGraph* was proposed [Ord+19]. In the following article [Ord+21], we test the capabilities of *SensoGraph*, analyzing a set of nine commercial chocolate chip cookies. The input contains data from 349 different participants, which is the largest amount of data used so far in projective mapping.

In order to provide an alternative to statistical methods, *SensoGraph* uses graph theory [Ord+19]. Samples become vertices of an undirected geometric graph. Using this framework, we compare two clustering modes: a proximity graph named Gabriel Graph, which conditions the existence of an edge between two vertices to a condition on their neighbourhood; and measuring the euclidean distance of the edges linking two vertices. Graphs can be translated into adjacency matrices, which can be added up, forming a global adjacency matrix, also called global similarity matrix. As a final step, the resulting similarity matrix can be used as an input for graph-drawing algorithms like Kamada-Kawai, yielding a resulting consensus graph.

The quality of the results is studied at the end of the article, showing a highly stable output with similar information compared to the statistical method. The scalability of *SensoGraph* is an important improvement, since it is linear on the number of participants. Moreover, the calculations performed by *SensoGraph* are independent for each participant, meaning that new data can be added at any time with minimal computational overhead.



Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Geometric and statistical techniques for projective mapping of chocolate chip cookies with a large number of consumers

David Orden^{a,*}, Encarnación Fernández-Fernández^b, Marino Tejedor-Romero^a,
Alejandra Martínez-Moraian^a

^a Departamento de Física y Matemáticas, Universidad de Alcalá, Campus Universitario, Ctra. Madrid-Barcelona, Km. 33.600, 28805 Alcalá de Henares, Spain

^b Área de Tecnología de Alimentos, E.T.S. de Ingenierías Agrarias, Universidad de Valladolid, Campus La Yutera, Avda. Madrid 50, 34004 Palencia, Spain



ARTICLE INFO

Keywords:

Projective mapping
Chocolate chip cookies
Consumers
Multiple factor analysis
SensoGraph
Stability

ABSTRACT

The so-called rapid sensory methods have proved to be useful for the sensory study of foods by different types of panels, from trained assessors to unexperienced consumers. Data from these methods have been traditionally analyzed using statistical techniques, with some recent works proposing the use of geometric techniques and graph theory. The present work aims to deepen this line of research introducing a new method, mixing tools from statistics and graph theory, for the analysis of data from Projective Mapping. In addition, a large number of $n = 349$ unexperienced consumers is considered for the first time in Projective Mapping, evaluating nine commercial chocolate chips cookies which include a blind duplicate of a multinational best-selling brand and seven private labels. The data obtained are processed using the standard statistical technique Multiple Factor Analysis (MFA), the recently appeared geometric method SensoGraph using Gabriel clustering, and the novel variant introduced here which is based on the pairwise distances between samples. All methods provide the same groups of samples, with the blind duplicates appearing close together. Finally, the stability of the results is studied using bootstrapping and the RV and Mantel coefficients. The results suggest that, even for unexperienced consumers, highly stable results can be achieved for MFA and SensoGraph when considering a large enough number of assessors, around 200 for the consensus map of MFA or the global similarity matrix of SensoGraph.

1. Introduction

Traditionally, the sensory evaluation of foods has been carried out using generic descriptive analysis with trained panels (Lawless & Heymann, 2010) and consumer sensory analysis with a 9-point hedonic scale, with different purposes and objectives (Varela & Ares, 2012). As an alternative to QDA and consumer sensory analysis with hedonic scale, a number of new sensory techniques have arisen in the last couple of decades (Varela & Ares, 2012, 2014; Valentin, Cholet, Nestrud, & Abdi, 2016). These methods, sometimes called rapid sensory methods as opposed to the fact that generic descriptive analysis is quite time consuming, have proved to be useful in order to obtain accurate and reliable information from consumers.

Among these rapid methods, the present work focuses on Projective Mapping (Risvik, McEwan, Colwill, Rogers, & Lyon, 1994), later used in Napping (Pagès, 2005). Some of the interesting characteristics of this method are being a holistic methodology, based on consumers' perception of the global similarities and differences among a set of samples, as well as allowing its use both with trained panellists and with

unexperienced consumers. Data from rapid methods are traditionally analyzed using statistical techniques, with multiple factor analysis (MFA) (Pagès, 2005) being the most common choice for Projective Mapping. Very recently, some works have also proposed the use of geometric techniques from graph theory (Orden, Fernández-Fernández, Rodríguez-Nogales, & Vila-Crespo, 2019; Lahne, 2020). A first contribution of the present work is to deepen in this recent line of research, introducing a substantial modification of the SensoGraph method proposed in (Orden et al., 2019) and further comparing the results and their stability between this new variant, the original SensoGraph, and MFA. Our hypothesis is that mixing geometric techniques from graph theory with common tools in statistics, like distances between samples or dendrograms, would provide further insight in sensory studies.

The number of assessors required to perform Projective Mapping differs depending upon the study, the products, and the level of expertise of the participants. Previous research has shown that the minimum number of consumers needed to obtain stable maps in Projective Mapping using MFA strongly depends on the number of samples and their degree of difference (Vidal et al., 2014; Vidal, Jaeger,

* Corresponding author.

E-mail address: david.orden@uah.es (D. Orden).

<https://doi.org/10.1016/j.foodqual.2020.104068>

Received 18 December 2019; Received in revised form 21 August 2020; Accepted 23 August 2020

Available online 27 August 2020

0950-3293/ © 2020 Elsevier Ltd. All rights reserved.

Antúñez, Giménez, & Ares, 2016). Although stable configurations might be reached by just 20 assessors, larger numbers are often required (Valentin et al., 2016).

Despite the simplicity of Projective Mapping makes it especially suitable to be used by unexperienced consumers, to the best of our knowledge no work seems to have carried out Projective Mapping with significantly more than 100 consumers, although other techniques like Sorting have been used with up to 389 consumers (Teillet, Schlich, Urbano, Cordelle, & Guichard, 2010). A second contribution of the present work is to perform a Projective Mapping study with a large number of consumers, in particular $n = 349$. This large number properly allows the use of resampling techniques in order to analyze the stability of both MFA and SensoGraph for a panel of unexperienced consumers. The data collected are made publicly available (Orden & Fernández-Fernández, 2020), seeking that the community can benefit from a dataset with such a large number of consumers.

As for the product under study, Projective Mapping has been previously used to analyze a variety of foods with different sensory complexities, from white wines (Pagès, 2005; Barton, Hayward, Richardson, & McSweeney, 2020) to cheeses (Barcenas, Elortondo, & Albisu, 2004; Nestrud & Lawless, 2010). For the present work we have chosen to study commercial chocolate chip cookies. Cookies are one of the most common snack foods (Gilbert, Miller, Olson, & St-Pierre, 2012) due to their general acceptability, convenience and shelf-life. In particular, chocolate chip cookies are present in most supermarket shelves, food stalls, and service stations. Their easy acquisition and low price help them become part of the diet, especially among young people (de Agricultura, 2019), which are the target of the present study. A Scopus search provided no previous works using Projective Mapping with chocolate chip cookies and a single one with commercial cookies (Tarrega, Marcano, & Fiszman, 2017). Thus, a third contribution of the present work is the analysis of commercial chocolate chip cookies, from a multinational brand and several private labels. Our hypothesis is that unexperienced consumers can provide consistent results about this popular type of cookies.

2. Material and methods

2.1. Samples

Nine commercial chocolate chip cookies were used in this study, bought at supermarkets in Palencia (Spain). Products differed in terms of brand, being or not a private label, manufacturer, and percentage of chocolate (Table 1). Blind duplicates were used within the product set, with samples 2 and 5 being the same product, since the ability of positioning close blind duplicates is widely considered an indicator of the reliability of the method and the accuracy of the panel (Moussaoui & Varela, 2010; Veinand, Godefroy, Adam, & Delarue, 2011; Savidan & Morris, 2015; Moelich, Muller, Joubert, Næs, & Kidd, 2017).

The chocolate chip cookies were presented in plastic cups, labelled with a three-digit random code, and served in randomized order following a balanced block experimental design. Due to the significant

differences in the size of the cookies and their external appearance, they were served as halves in order to minimize the possibility of the consumers recognizing them. To make sure that the assessors could re-test several times if needed, labelled plates with extra samples were at their disposal. Water was also provided to all consumers to rinse between samples.

2.2. Consumers

A total of three hundred forty-nine ($n = 349$) consumers participated in this study, recruited among seven educational centers and two university fairs along the academic years 2017–18 and 2018–19; more details are provided in the dataset description (Orden & Fernández-Fernández, 2020). The participants, of which 53% were female and 47% male, ranged between 14 and 30 years old. All the participants agreed to take part in this study and no identifiable or sensitive information was collected.

2.3. Projective mapping

Prior to starting the tasting session, all participants were sensitized about the importance of sensory analysis by a visual presentation. An aim underlying this study was to promote the sensory analysis among young people, at different high schools and during the open days of the University of Valladolid. Subsequently, the basis of Projective Mapping was explained to the participants, using an illustration which showed cookies of different shapes (rectangular and round) and colors (cream and brown).

After the explanation of the technique, the participants received an A2 (60 × 40 cm) sheet of paper to allocate the samples. Samples were to be placed close to each other if, according to the assessor's own criteria, they seemed sensorially similar and vice versa, i.e., two chocolate chip cookies were to be distant from one another if they seemed different. The participants had to observe, smell, and taste the chocolate chip cookies, and then position the samples on the A2 sheet, trying to use as much of the tablecloth as possible. Once they had decided on the positioning, they were asked to write the codes on the sheet. The dataset with the x- and y-coordinates of chocolate chip cookies from the individual perceptual spaces is made publicly available (Orden & Fernández-Fernández, 2020), for the sake of facilitating replicability in research.

2.4. Data analysis with existing methods

2.4.1. Statistical analysis

The data obtained from Projective Mapping (Risvik et al., 1994) was then analyzed by MFA (Pagès, 2005), using the R language (R Development Core Team, 2007) and the FactoMineR package (Lê, Josse, & Husson, 2008). Confidence ellipses were constructed using truncated total bootstrapping (Cadoret & Husson, 2013) with the SenoMineR package (Lê & Husson, 2008).

Table 1
Chocolate chip cookies samples information.

Sample	Brand	Private label of supermarket	Manufacturer	Percentage of Chocolate
1	Hacendado	Mercadona	Grupo Siro	37% chocolate chips
2	Chips-Ahoy		Mondelez España Commercial, S.L.	25.6% chocolate chips
3	Carrefour	Carrefour	Aurly S.L.	37% chocolate chips
4	Grandino	Lidl	Übach-Palenberg	29% chocolate chips and 11% milk chocolate chips
5	Chips-Ahoy		Mondelez España Commercial, S.L.	25.6% chocolate chips
6	Alteza	Lupa	Galletas Gullón S.A.	25% chocolate chips
7	American Cookies	Aldi	Banketbakkerij Merba B.V.	29% chocolate chips and 11% milk chocolate chips
8	Dia	Dia	Don Cake S.A.	26.7% chocolate chips and 10.3% milk chocolate chips
9	Ifa Eliges	Gadis	Galletas Gullón S.A.	25% chocolate chips

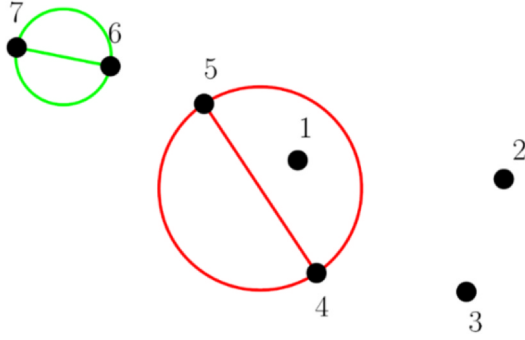


Fig. 1. Clustering with the Gabriel graph. The connection 6–7 will be drawn, since no third sample lies inside the circle having that connection as diameter. On the contrary, the connection 4–5 will not be drawn, because sample 1 lies inside the corresponding circle.

2.4.2. Geometric analysis using SensoGraph with a clustering method

The x- and y-coordinates from the tablecloths were imported and analyzed using a web app (Orden & Tejedor-Romero, 2019) implementation of the SensoGraph method introduced in (Orden et al., 2019). The first step there was to perform a clustering on each tablecloth using the Gabriel graph (Gabriel & Sokal, 1969), a tool from Computational Geometry. With this technique, two samples become connected if, and only if, there is no third sample contained in the circle having that potential connection as diameter. See Fig. 1.

The aim of this clustering was to connect some pairs of samples at each tablecloth, so that a global similarity matrix could be constructed by counting, for each pair of samples, in how many tablecloths they became connected.

In other words, for each tablecloth the clustering induced a matrix with entries in the set {0,1} standing, respectively, for the corresponding pair of samples being connected or not by the clustering. Then, the global similarity matrix was just the addition of the particular similarity matrices from each tablecloth. See Fig. 2 for an illustration.

The second step in (Orden et al., 2019) was the use of the Kamada-Kawai force-directed algorithm from graph drawing (Kamada & Kawai, 1989) to obtain a positioning of the samples (consensus graphic) based on the entries of the global similarity matrix.

Interested readers can follow the descriptions in (Orden et al., 2019) to implement the SensoGraph method in R, using the command gg in the package cccd (Marchette, 2015) for the Gabriel graph and the command layout_with_kk of the package igraph (Csardi, 2015) for the Kamada-Kawai graph drawing algorithm. The corresponding author can also be contacted for further details.

Several improvements were implemented with respect to the software used in (Orden et al., 2019) and the features of the above-mentioned commands in R. First, a color code from red (smallest) to green (largest) was incorporated to both the connections between samples in the consensus graphic and the global similarity matrix, to illustrate the strength of those connections following the lines of recent successful visualization tools (YouGov, 2019). Second, a dendrogram for the data of the global similarity matrix was obtained using hierarchical clustering (Härdle & Simar, 2007) and the matrix was then rearranged

according to the order in the dendrogram, for the sake of an easier visualization of the groups of samples.

In addition, the possibility of displaying only the most relevant connections was implemented. The strength of a particular connection was normalized to the interval between the smallest and largest strengths in the global similarity matrix, according to the following formula

$$\text{normalized strength} = \frac{\text{current strength} - \text{smallest strength}}{\text{largest strength} - \text{smallest strength}}$$

Then, these normalized strengths were grouped into deciles, allowing to display only the 10-k% of largest normalized strengths, for k an integer from 1 to 10.

2.5. Data analysis with a new method

The present work introduces and tests an alternative way of obtaining the global similarity matrix. As in the original SensoGraph proposal, the global matrix will also be the sum of particular similarity matrices for each tablecloth, but the difference is that the values in these particular matrices now lie in the interval [0,1]. See Fig. 3 for an illustration.

Specifically, the values of the particular similarity matrices correspond to the distances between the pairs of samples, normalized to the interval [0,1] as follows: The interval between the smallest and the largest distances obtained is linearly mapped to a similarity in the interval [0,1] in an inverse way, according to the formula

$$\text{similarity} = 1 - \frac{\text{distance} - \text{smallest distance}}{\text{largest distance} - \text{smallest distance}}$$

which assigns similarity 1 to the smallest distance, 0 to the largest distance, smaller similarities (between 0 and 1) to the larger distances, and vice versa.

This approach aims to capture the essence of Projective Mapping, which asks the assessors to position closer those samples perceived as more similar and vice versa, by assigning larger similarities to (hence, considering more important) those pairs of samples positioned closer, and vice versa. In order to emphasize the importance of those samples which were more clearly positioned closer or further, the linear mapping mentioned above was tuned using the following formula, which depends on a parameter $p \geq 1$

$$\text{tuned similarity} = \begin{cases} 2^{p-1} \cdot (\text{similarity})^p & , \text{when similarity} < 1/2 \\ 1 - 2^{p-1} \cdot |\text{similarity} - 1|^p & , \text{when similarity} \geq 1/2 \end{cases}$$

Fig. 4 shows the effect of this function for several values of p. The value $p = 1$ gives the aforementioned linear mapping, while increasing the value of p results in the graph combing towards an S-shape, hence emphasizing the effect of extreme values corresponding to more clearly positioned samples. After checking different values of p between 1 and 3 for the function of tuned similarity, obtaining analogous results, a compromise value of $p = 2$ was chosen for the sake of emphasizing extreme values without an excessive distortion of the inputs, see again Fig. 4.

This new method was also implemented in the web app (Orden &

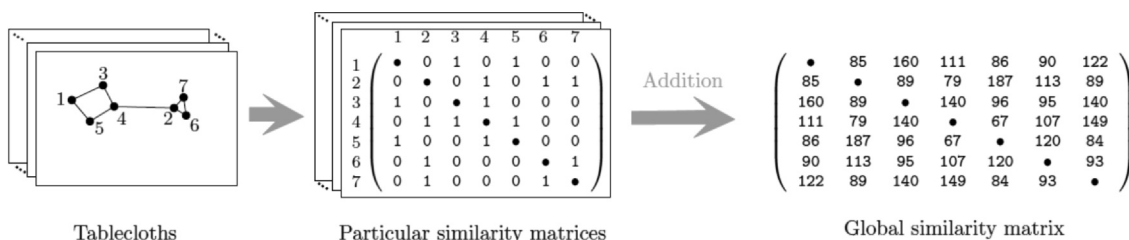


Fig. 2. Illustration of the process for SensoGraph with Gabriel (Orden et al., 2019).

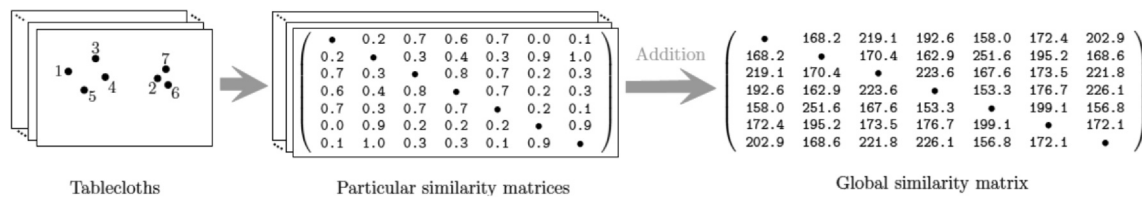


Fig. 3. Illustration of the process for SensoGraph with distances.

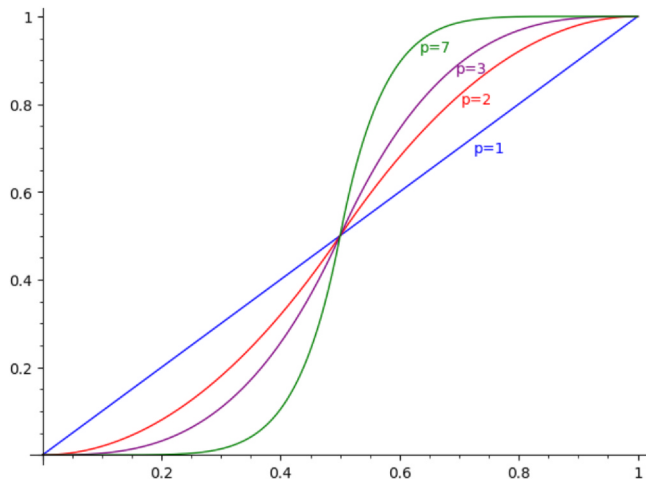


Fig. 4. Effect of linear ($p = 1$) and non-linear mappings ($p > 1$) over the interval $[0,1]$.

Tejedor-Romero, 2019), using Python as backend language on the server side, Flask as microframework, and MongoDB to manage the databases, as well as JavaScript, HTML, and CSS on the user side. Again, replicability of the results can be ensured by alternative implementations using the techniques and commands detailed above. Interested readers can also contact the corresponding author for further details.

2.6. Stability of the results

The stability of the results obtained was analyzed using bootstrapping resampling in order to simulate repetition of the experiments (Shao & Tu, 1995), as used by Vidal et al. (2014) for the study of the stability of sample configurations from projective mapping. For each value of $m = 10, 20, 30, \dots, n$, subsets of m assessors were randomly drawn with replacement from the original data set. As in previous works (Faye et al., 2006; Blancher, Clavier, Egoroff, Duineveld, & Parcon, 2012), a collection of 100 subsets was generated for each value of m .

For the study of the stability of MFA results three analyses were performed, using independent bootstrapping resamplings: With the first two dimensions of the MFA and with the first four dimensions as in the stability study by Vidal et al. (2014) and, in addition, with eight dimensions for better comparison with the results for SensoGraph.

In order to measure the agreement between the MFA consensus map for each subset and that of the original panel, the RV coefficient was computed (Escoufier, 1973; Tomic, Berget, & Næs, 2015; Josse & Holmes, 2016). This is a popular similarity measure between point configurations or matrices, whose values range between 0 and 1 and for which, the more similar two items are, the higher is the corresponding RV coefficient.

The RV coefficient was computed using the FactoMineR (Lê et al., 2008) function `coeffRV`. The mean and the standard deviation for RV coefficients of subsets of the same size were then obtained using the commands `mean` and `sd` of the R language (R Development Core Team,

2007).

For the study of the stability of SensoGraph results, the only meaningful analysis is considering the global similarity matrices. A graph is defined by a set of vertices and a set of pairs (i,j) of vertices (Gross, Yellen, & Zhang, 2013), which are encoded as the set of entries (i,j) of a matrix. The two-dimensional representation of a global similarity matrix is a plotting artifact based on the graph-drawing algorithm considered. Furthermore, the random nature of the seeds chosen by the Kamada-Kawai algorithm used by SensoGraph could result in different coordinates in the consensus graph.

In order to measure the agreement between the SensoGraph similarity matrix for each subset and that of the original panel, the Mantel coefficient was computed (Mantel, 1967), which evaluates the similarity between two configurations by measuring the correlation between two matrices of distances between the samples (Abdi, 2010). The Mantel coefficient is the usual Pearson correlation coefficient based on two vectors of size $n(n-1)/2$ (n being the number of products) containing the off-diagonal elements of the two dissimilarity matrices being compared. Its values range between -1 and 1 and, the more similar two items are, the higher is the corresponding Mantel coefficient. Widely used in ecology, the Mantel coefficient has also been used in sensory analysis, e.g., by Blancher et al. (2012) to investigate the stability of sorting maps.

The mantel coefficient was computed using the Python function `skbio.stats.distance.mantel` (The scikit-bio development team, 2020) using Pearson's product-moment correlation coefficient. The mean and the standard deviation for Mantel coefficients of subsets of the same size were then obtained using the commands `mean` and `dataset.mean` and `dataset.std` of the pandas library (The Pandas Development Team, 2020).

3. Results and discussion

3.1. Results for MFA

The data obtained from the Projective Mapping tablecloths were exported as a CSV file and were processed using MFA as mentioned in Section 2.4.1. The computation took an average of 90.2 s, with a standard deviation of 11.7 s, measured over 5 independent runs on a computer with an Intel(R) Core(TM)2 Duo CPU @2.13 GHz with 4 GB of RAM.

The consensus representation of the similarities and differences among samples is shown in Fig. 5. In this consensus graphic of MFA, the first two dimensions accounted for 41.12% of the explained variance (25.64% Dim1 and 15.48% Dim2). Low percentages of explained variance have been observed for panels composed by unexperienced consumers (Nestrud & Lawless, 2010), with some works reporting such panels not positioning close together blind duplicate samples (Nestrud & Lawless, 2008).

In our case, the panel of unexperienced consumers was able to position the blind duplicate samples 2 and 5, from the same brand (Chips-Ahoy), close together in the lower-right quadrant. Their confidence ellipses overlapping means that no significant differences were perceived and, therefore, the consumers were capable to detect their similarity.

In addition to that, samples 1 (Hacendado), 3 (Carrefour), 4

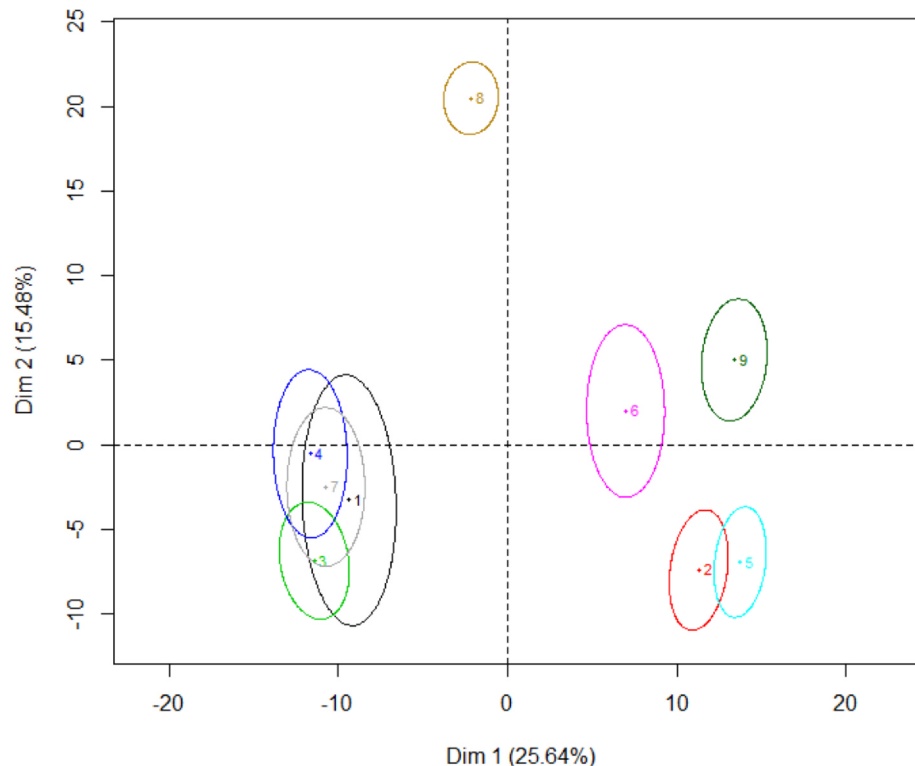


Fig. 5. Consensus plot from MFA, dimensions 1 and 2.

(Grandino), and 7 (American Cookies) became positioned in the lower-left quadrant of the perceptual space, with their confidence ellipses overlapping indicating that consumers did not perceive statistically significant differences between them. These were the cookies with the highest percentage of chocolate: Samples 1 (Hacendado) and 3 (Carrefour) had a 37% of chocolate chips, while samples 4 (Grandino) and 7 (American Cookies) had a 29% of chocolate chips plus an 11% of milk chocolate chips. Recall Table 1.

Further, sample 6 (Alteza) and sample 9 (Ifa Eliges) appear separated, with their ellipses not overlapping but not far to each other and to samples 2 and 5. Probably, this can be explained because the samples 6 and 9 have a percentage of chocolate similar to samples 2 and 5 (25% and 25.6% of chocolate chips, respectively, see Table 1). Samples 6 and 9 are from a private label, but produced by the same manufacturer, while the duplicate samples 2 and 5 are from a best-selling multinational brand (see Table 1).

Finally, sample 8 (Dia), was isolated at the upper-left quadrant despite its proportion of chocolate chips being similar to others (26.7% of chocolate chips and 10.3% of milk chocolate chips, see Table 1). A possible explanation is that this sample appears as the farthest one to samples 2 and 5, those from a best-selling multinational brand, so the private label cookie 8 was perceived as the least similar to a best-selling cookie.

The above discussion is to be considered taking into account that, due to the lack of batch tracking, changes in product recipes might have arisen over the two years of data collection.

3.2. Results for SensoGraph with Gabriel clustering

The data from the Projective Mapping tablecloths was uploaded to the SensoGraph software (Orden & Tejedor-Romero, 2019), which processed them as described in Section 2.4.2. The computation took less than one second with the same computer mentioned in Section 3.1.

Fig. 6 shows the consensus graphic obtained, with all the connections between samples, obtained using SensoGraph with the Gabriel clustering as in (Orden et al., 2019), see Section 2.4.2. Fig. 7

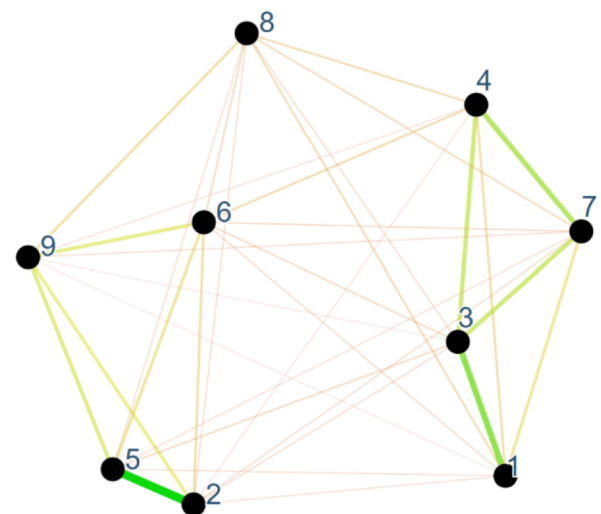


Fig. 6. Consensus plot obtained using SensoGraph with Gabriel (Orden et al., 2019).

shows only the 60% of most relevant connections, while Fig. 8 shows the dendrogram (left) obtained by hierarchical clustering from the global similarity matrix with the strengths of those connections (right). For the consensus plots and the global similarity matrix, a color code from red (smallest strength) to green (largest strength) has been used for the sake of an easier visualization.

These graphics show that the positioning of the samples provided by SensoGraph is similar to that in the consensus map given by MFA, the RV coefficient between the point configurations in Figs. 5 and 6 being 0.8785. Groups 2–5–6–9, 1–3–4–7, and 8 can be identified in Figs. 7 and 8, with the global similarity matrix in Fig. 8 (right) easily showing that the connection 2–5 is the strongest (greenest) one, corresponding to samples 2 and 5 being the blind duplicates. It is interesting to note that SensoGraph identifies these blind duplicates as the most similar

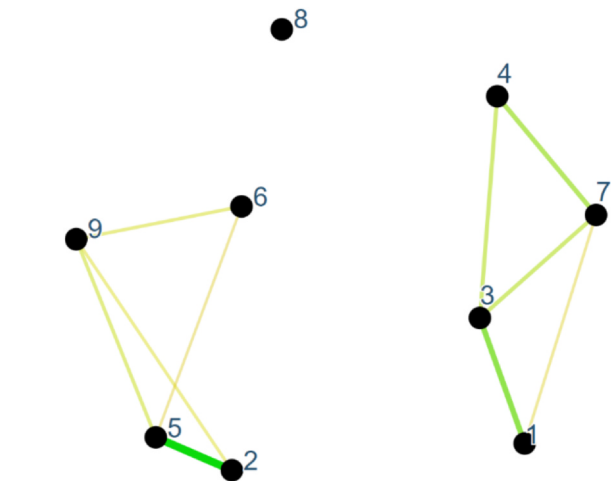


Fig. 7. Consensus plot from SensoGraph with Gabriel, showing only the 60% of most relevant edges.

samples, both by the consensus map in Fig. 6 and by the global similarity matrix in Fig. 8, while in the MFA graphic for the first two dimensions (Fig. 5) the pair 1–7 is instead the closest one (although the confidence ellipses for 2–5 do overlap, as discussed above). Nevertheless, note that the value of the connection 2–5 in the SensoGraph similarity matrix is just 187 out of the 349 consumers considered.

In addition, SensoGraph shows samples 6 and 9 appearing in the same group as the aforementioned samples 2 and 5, although the strength of the other connections in this group is quite smaller than that of the connection 2–5. Samples in this group share a similar percentage of chocolate chips since, as commented above, samples 6 (Alteza) and 9 (Ifa Eliges) have a 25% of chocolate chips, and samples 2 and 5 (Chips-Ahoy) have a 25.6% of chocolate chips (Table 1).

A second group observed in the SensoGraph graphic is composed by samples 1–3–4–7, with the corresponding ellipses overlapping in MFA as well. The connection 1–3 is the second strongest one and the connection 4–7 is the third strongest one, while connections 1–4 and 1–7 are not so strong, and the connections 3–4 and 3–7 have an intermediate strength. This group can be explained because the samples 1 (Hacendado) and 3 (Carrefour) have the same percentage of chocolate chips, 37%, while the sample 4 (Grandino) and the sample 7 (American Cookies) also have both the same percentage of chocolate (29% of chocolate chips plus 11% of milk chocolate chips).

Finally, the sample 8 (Dia) being isolated in MFA coincides with its connections to all other samples being weak in SensoGraph, and with this sample being actually isolated when showing only the 60% of most relevant edges (Fig. 7). As happened for MFA, the sample 8 appears as the farthest one to samples 2 and 5 also in SensoGraph, both for the consensus map (Fig. 6) and for the global similarity matrix (Fig. 8,

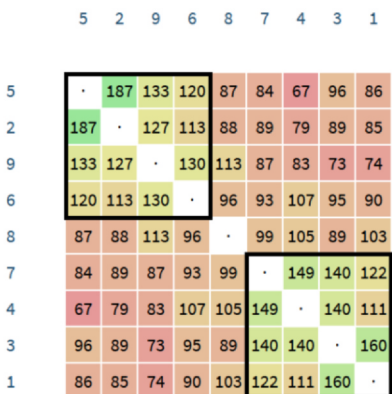
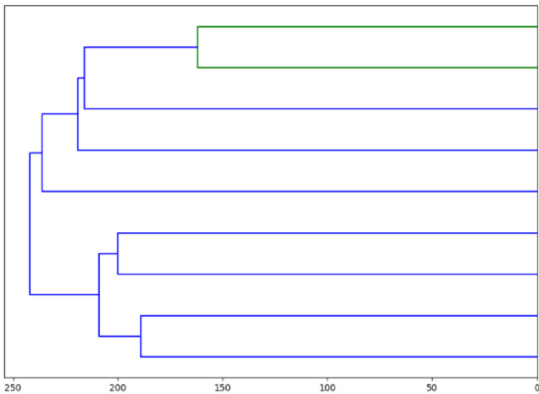


Fig. 8. Dendrogram (left) and global similarity matrix (right) for SensoGraph with Gabriel. The matrix is rearranged according to the result of the dendrogram, so that the groups obtained appear as submatrices with similar colors (framed). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

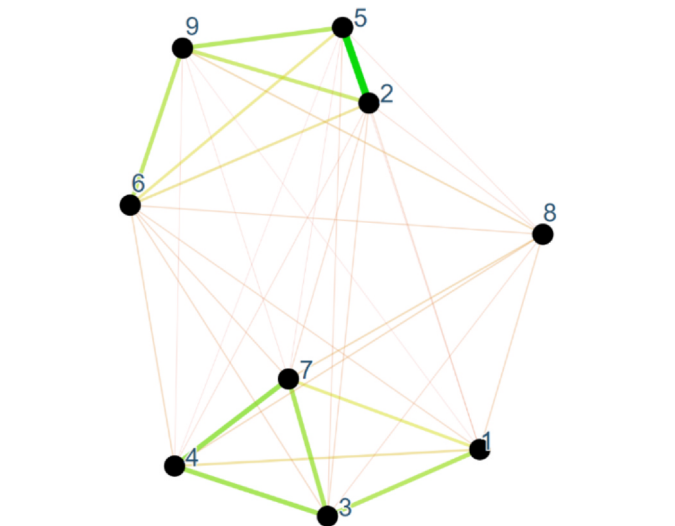


Fig. 9. Consensus plot obtained using SensoGraph with distances, as introduced in Section 2.5.

right).

3.3. Results for SensoGraph with distances

In this case, the data was uploaded to the software by Orden and Tejedor-Romero (2019) and the option of using SensoGraph with distances was selected, in order to process the data as described in Section 2.5. The computation took less than one second with the same computer mentioned in Section 3.1.

Fig. 9 shows the consensus graphic obtained, with all the connections between samples, using SensoGraph with distances, as introduced in Section 2.5. Fig. 10 shows only the 60% of most relevant connections. Finally, Fig. 11 depicts the dendrogram (left) obtained from the global similarity matrix (right). Again, the color code from red to green has been used for the consensus graphics and the similarity matrix.

The consensus map obtained (Fig. 9) is similar to those given by the previous methods, the RV coefficient with MFA for the first two dimensions (Fig. 5) being 0.7407 and that with SensoGraph with Gabriel (Fig. 6) being 0.6390. Both the consensus maps (Figs. 5, 6, and 9) and the dendrograms and global similarity matrices (Figs. 8 and 11) show the same groups, being slightly more clear for SensoGraph with distances than for the variant with Gabriel (Figs. 7 and 10, also Figs. 8 and 11).

3.4. Stability of the results

As detailed in Section 2.6, bootstrapping was performed to analyze

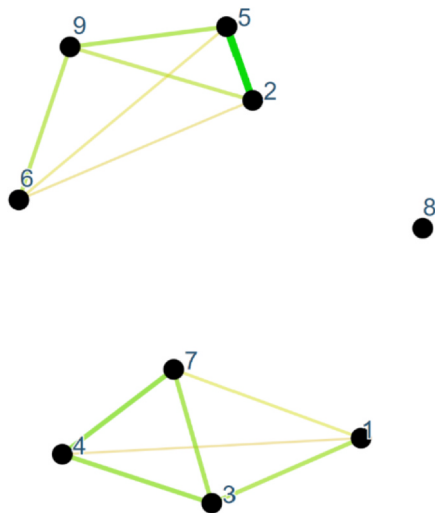


Fig. 10. Consensus plot from SensoGraph with distances, showing only the 60% of most relevant edges.

the stability of the results given in the previous subsections by the three methods considered.

Figs. 12–16 show, for different cases, the evolution of the RV and Mantel coefficients between virtual panels, composed by subsets of consumers randomly drawn with replacement, and the true panel. The vertical axis corresponds to the average RV or Mantel coefficient, incorporating standard deviations as vertical bars, while the horizontal axis corresponds to the number of consumers in the virtual panel. As expected, in all the cases increasing the number of consumers leads to an increase of the RV or Mantel coefficients and a decrease of the standard deviations.

Previous works (Faye et al., 2006, Blancher et al., 2012; Vidal et al., 2014) have considered different values of the RV coefficient as the threshold above which the results are considered to be stable, the most restrictive one being the value 0.95 proposed by Blancher et al. (2012), which is depicted as a red horizontal dashed line. There is no agreement in the literature about which value of the RV coefficient indicates a good agreement, with Vidal et al. (2014) reporting works which consider values that range between 0.65 and 0.95. For the Mantel coefficient there is no standard threshold value as well, although Blancher et al. (2012) observed a strong linear relationship between both types of coefficient, with the Mantel coefficient tending to provide slightly lower values than the RV coefficient.

For the case of MFA, considering the first four dimensions instead of the first two leads to better results, with higher initial RV coefficients and smaller standard deviations. Vidal et al. (2014) observed similar results where, their work considering up to 100 consumers, some panels did not reach the 0.95 RV coefficient suggested by Blancher et al.

(2012). The results we obtain show that for this panel of unexperienced consumers with the first two dimensions accounting for only 41.12% of the explained variance, the 0.95 threshold is achieved for around 200 consumers (Fig. 12). When considering the first four dimensions, accounting for 65.95% of the explained variance, the 0.95 threshold is achieved for around 150 consumers (Fig. 13). Considering the eight dimensions, which obviously account for the 100% of the explained variance, allows to achieve the 0.95 threshold for 40 consumers (Fig. 14).

It should be noticed that Næs, Berget, Liland, Ares, and Varela (2017) analyzed the corpus of 46 publications in Food Quality and Preference and Food Research International dealing with projective mapping until then, concluding that most of the papers considered only the first two dimensions of the MFA, with just a few of them (6 out of 46) going further and using up to the first four dimensions. To the best of our knowledge, no previous work used the full (eight) dimensions of MFA to study the stability of the samples configuration.

As for SensoGraph with Gabriel, the results in Fig. 15 show smaller deviations than those for MFA with the first two dimensions (Fig. 12), although more consumers are needed to achieve the 0.95 threshold, around 300 consumers. This could be explained because, as observed by Blancher et al. (2012), the Mantel coefficient tends to be slightly smaller than the RV coefficient. Actually, in this case, considering 200 consumers achieves a 0.92 threshold.

It is worth noticing that, although the dimensionality of the 9x9 global similarity matrix is eight (it defines nine points in nine dimensions but the fact that entries (i,i) are zero implies that those points do actually lie on an 8-dimensional hyperplane), the global similarity matrix can also be depicted in a single 2-dimensional graphic (Fig. 8, right).

Finally, SensoGraph with distances leads to the results depicted in Fig. 16, which show higher Mantel coefficients and smaller standard deviations than the version using Gabriel. In particular, the 0.95 threshold is achieved for around 200 consumers, like for MFA with the first two dimensions (Fig. 12), but with smaller deviations. Further studies with different panels and products, as in (Vidal et al., 2014), should be performed in order to confirm this behavior.

4. Conclusions

This work used Projective Mapping for the evaluation of commercial chocolate chip cookies by a large number of $n = 349$ unexperienced consumers, analyzing the data with statistical (MFA), geometric (SensoGraph-Gabriel) and mixed (SensoGraph-distances) methods. All of them provided the same groups of samples, with the two blind duplicates being positioned close together. The identification of these duplicates was clearer for the geometric and mixed techniques than for the MFA consensus map, where a different pair of samples appeared as the closest one.

The stability of the results was studied with bootstrapping

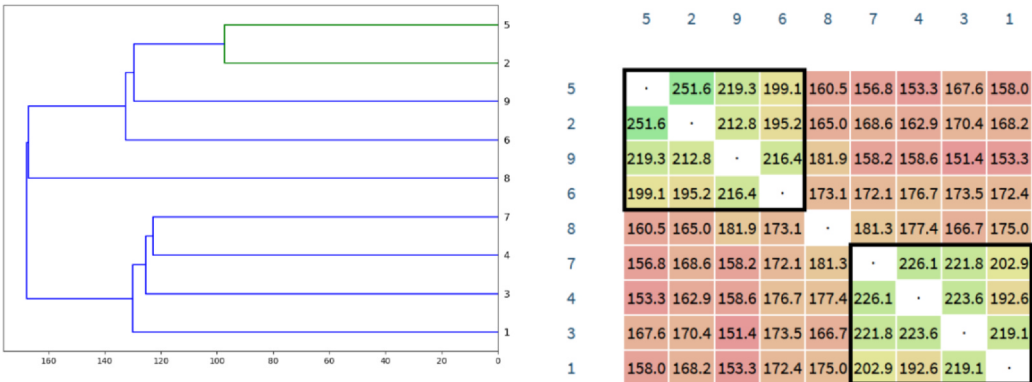


Fig. 11. Dendrogram (left) and global similarity matrix (right) for SensoGraph with distances. The matrix is rearranged according to the result of the dendrogram, so that the groups obtained appear as submatrices with similar colors (framed). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

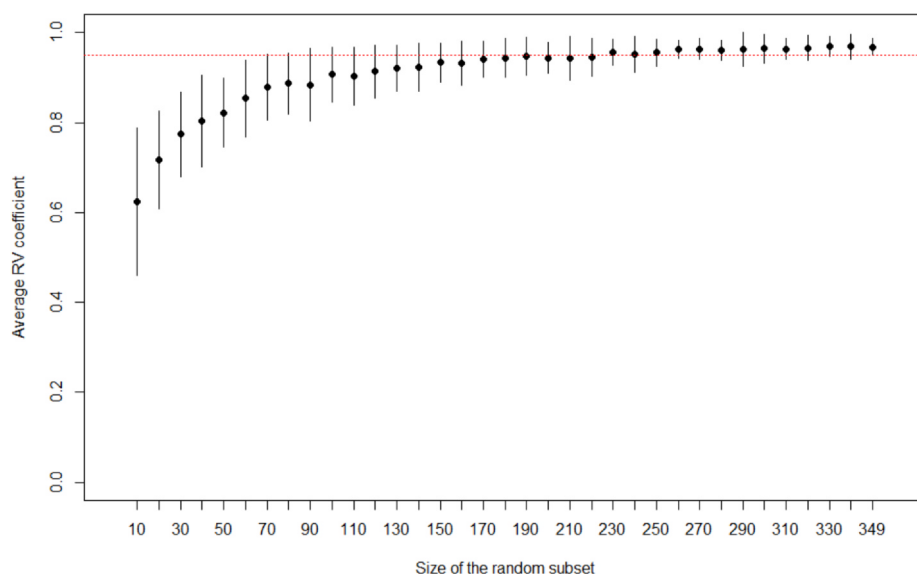


Fig. 12. Evolution of the RV coefficient between subsets of consumers randomly drawn with replacement and the whole panel, for the first two dimensions of MFA (Fig. 5). The red horizontal dashed line corresponds to a 0.95 threshold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

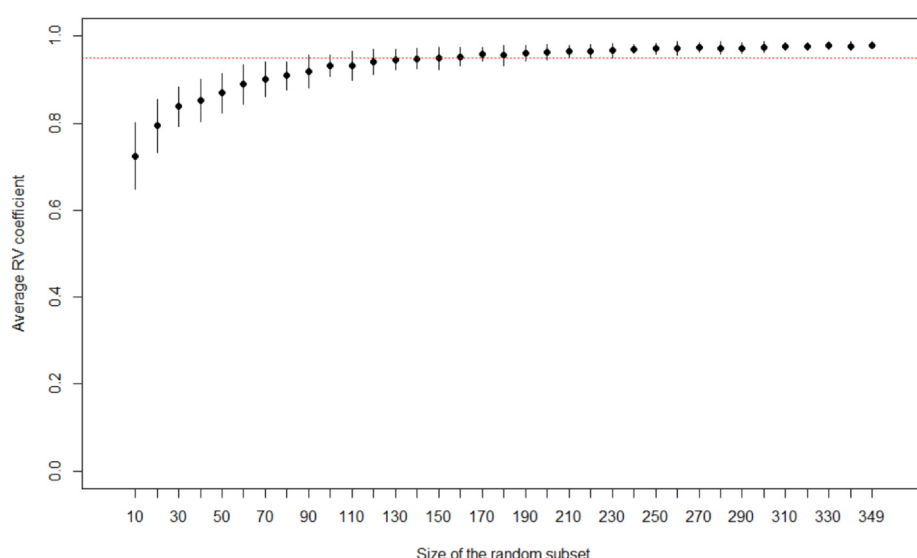


Fig. 13. Evolution of the RV coefficient between subsets of consumers randomly drawn with replacement and the whole panel, for the first four dimensions of MFA. The red horizontal dashed line corresponds to a 0.95 threshold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

resampling, randomly drawing 100 subsets of $m = 10, 20, 30, \dots, n$ assessors from the original data set and measuring the agreement with the original panel by the RV coefficient for MFA and the Mantel coefficient for SensoGraph. MFA achieved the highly restrictive RV 0.95 stability threshold for around 200 consumers when using the first two dimensions, for around 150 consumers when considering the first four dimensions, and for 40 consumers when all the eight dimensions are considered. For SensoGraph with Gabriel, the Mantel 0.95 stability threshold was achieved around 300 consumers, while SensoGraph with distances led to values beyond the Mantel 0.95 threshold for around 200 consumers. These values are to be considered taking into account that Mantel coefficients have been previously observed to be slightly smaller than RV coefficients.

Further research, with different panels and products, would be needed in order to confirm these behaviors which suggest that, on one hand, global similarity matrices are useful for Projective Mapping data analysis and, on the other hand, graph drawing techniques provide reliable consensus maps.

CRedit authorship contribution statement

David Orden: Conceptualization, Methodology, Validation,

Visualization, Supervision, Writing - original draft. **Encarnación Fernández-Fernández:** Validation, Formal analysis, Investigation, Writing - review & editing, Supervision. **Marino Tejedor-Romero:** Methodology, Software, Formal analysis, Visualization. **Alejandra Martínez-Moraian:** Formal analysis, Data curation, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

David Orden has been partially supported by Project CCG19/CC-035 (GINSENG) of the University of Alcalá, by Project MTM2017-83750-P of the Spanish Ministry of Science (AEI/FEDER, UE), by Project PID2019-104129GB-I00 of the Spanish Ministry of Science and Innovation, and by the European Union H2020-MSCA-RISE project 734922 – CONNECT. Encarnación Fernández-Fernández has been partially supported by Project CCG19/CC-035 (GINSENG) of the University of Alcalá and by Project PID2019-104129GB-I00 of the Spanish

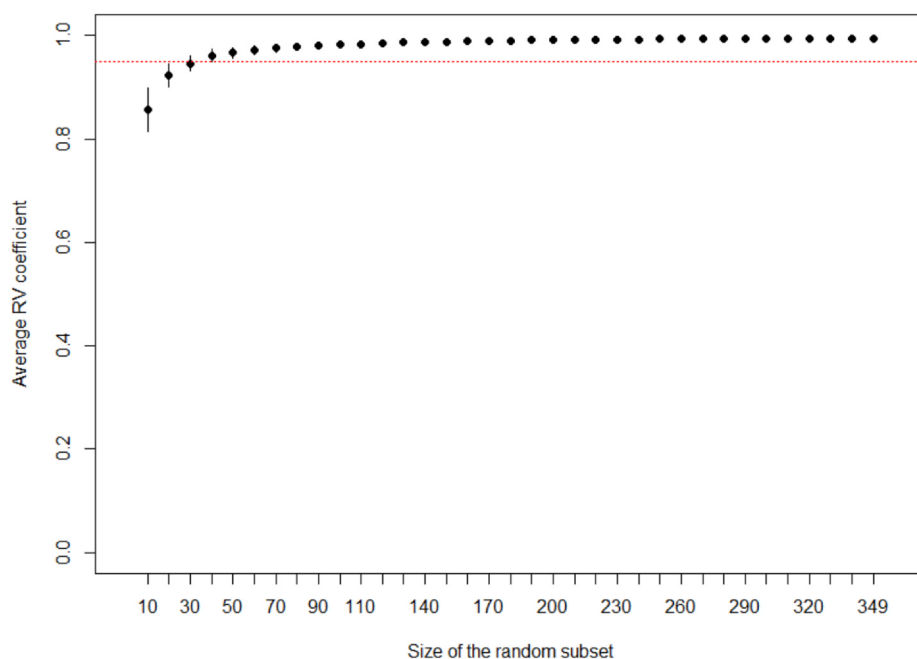


Fig. 14. Evolution of the RV coefficient between subsets of consumers randomly drawn with replacement and the whole panel, for the eight dimensions of MFA. The red horizontal dashed line corresponds to a 0.95 threshold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

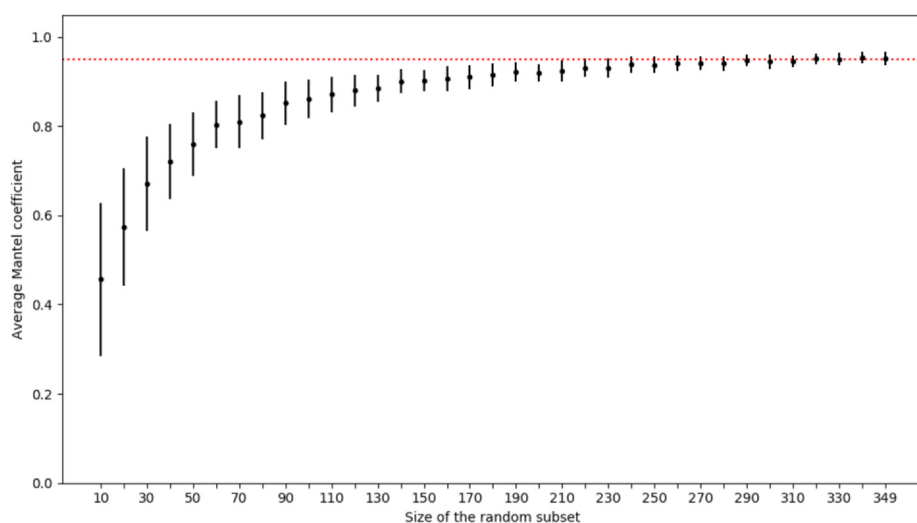


Fig. 15. Evolution of the Mantel coefficient between subsets of consumers randomly drawn with replacement and the whole panel, for the SensoGraph with Gabriel global similarity matrix (Fig. 8). The red horizontal dashed line corresponds to a 0.95 threshold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

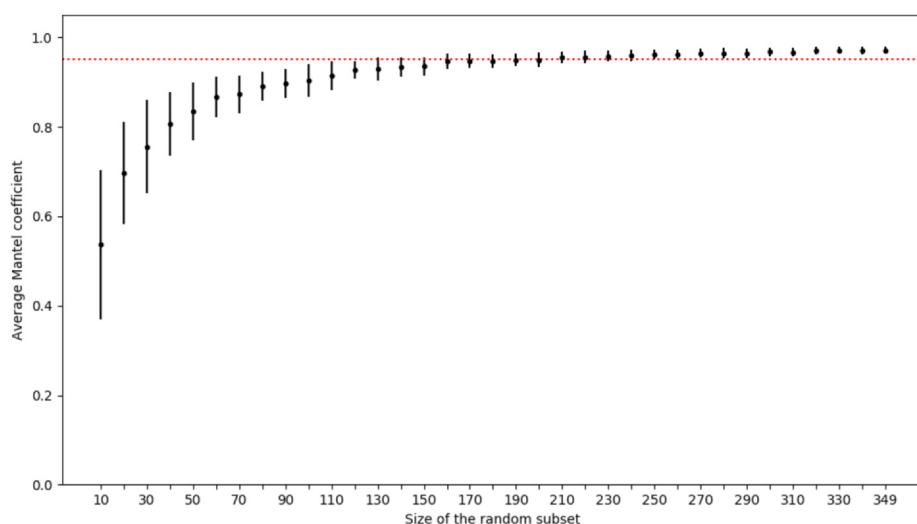


Fig. 16. Evolution of the Mantel coefficient between subsets of consumers randomly drawn with replacement and the whole panel, for the SensoGraph with distances global similarity matrix (Fig. 11). The red horizontal dashed line corresponds to a 0.95 threshold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Ministry of Science and Innovation. Marino Tejedor-Romero is funded by the University of Alcalá, through the program Ayudas de iniciación en la actividad investigadora. Alejandra Martínez-Moraian is funded by the predoctoral contract PRE2018-085668 of the Spanish Ministry of Science, Innovation, and Universities, and partially supported by Project PID2019-104129GB-I00 of the Spanish Ministry of Science and Innovation. Part of her work on this paper was made during a stay at the University of Alcalá while she was with the Departamento de Matemáticas, Estadística y Computación at the University of Cantabria.

References

- Abdi, H. (2010). Congruence: congruence coefficient, RV-coefficient and Mantel coefficient. In N. J. Salkind, D. M. Dougherty, & B. Frey (Eds.). *Encyclopedia of research design* (pp. 222–229). Thousand Oaks, CA: Sage.
- Barceñas, P., Elortondo, F. P., & Albiu, M. (2004). Projective mapping in sensory analysis of ewes milk cheeses: A study on consumers and trained panel performance. *Food Research International*, 37(7), 723–729. <https://doi.org/10.1016/j.foodres.2004.02.015>.
- Barton, A., Hayward, L., Richardson, C. D., & McSweeney, M. B. (2020). Use of different panellists (experienced, trained, consumers and experts) and the projective mapping task to evaluate white wine. *Food Quality and Preference*, 83, Article 103900. <https://doi.org/10.1016/j.foodqual.2020.103900>.
- Blancher, G., Clavier, B., Egoroff, C., Duineveld, K., & Parcon, J. (2012). A method to investigate the stability of a sorting map. *Food Quality and Preference*, 23, 36–43. <https://doi.org/10.1016/j.foodqual.2011.06.010>.
- Cadoret, M., & Husson, F. (2013). Construction and evaluation of confidence ellipses applied at sensory data. *Food Quality and Preference*, 28(1), 106–115. <https://doi.org/10.1016/j.foodqual.2012.09.005>.
- Csardi, G. (2015). igraph. <http://igraph.org/r/>. Accessed: 2019-12-18.
- Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, 29(4), 751–760.
- Faye, P., Brémaud, D., Teillet, E., Courcoux, P., Giboreau, A., & Nicod, H. (2006). An alternative to external preference mapping based on consumer perceptive mapping. *Food Quality and Preference*, 17(7–8), 604–614. <https://doi.org/10.1016/j.foodqual.2006.05.006>.
- Gabriel, K. R., & Sokal, R. R. (1969). A new statistical approach to geographic variation analysis. *Systematic Biology*, 18(3), 259–278. <https://doi.org/10.2307/2412323>.
- Gilbert, J. A., Miller, D., Olson, S., & St-Pierre, S. (2012). After-school snack intake among Canadian children and adolescents. *Canadian Journal of Public Health*, 103(6), e448–e452. <https://doi.org/10.1007/BF03405636>.
- Gross, J. L., Yellen, J., & Zhang, P. (2013). *Handbook of graph theory*. Chapman and Hall/CRC ISBN 9781439880180.
- Härdle, W., & Simar, L. (2007). *Applied multivariate statistical analysis* (pp. 1051–8215). Berlin: Springer. <https://doi.org/10.1007/978-3-662-45171-7>.
- Josse, J., & Holmes, S. (2016). Measuring multivariate association and beyond. *Statistics surveys*, 10, 132. <https://doi.org/10.1214/16-SS116>.
- Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1), 7–15. [https://doi.org/10.1016/0020-0190\(89\)90102-6](https://doi.org/10.1016/0020-0190(89)90102-6).
- Lahne, J. (2020). Sorting Backbone Analysis: A network-based method of extracting key actionable information from free-sorting task results. *Food Quality and Preference*, 103870. <https://doi.org/10.1016/j.foodqual.2020.103870>.
- Lawless, H. T., & Heymann, H. (2010). Sensory evaluation of food: Principles and practices. *Springer Science and Business Media*. <https://doi.org/10.1007/978-1-4419-6488-5>.
- Lê, S., & Husson, F. (2008). SensoMineR: A package for sensory data analysis. *Journal of Sensory Studies*, 23(1), 14–25. <https://doi.org/10.1111/j.1745-459X.2007.00137.x>.
- Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1–18. <https://doi.org/10.18637/jss.v025.i01>.
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27, 209–220.
- Marchette, D. J. (2015). ccc, Class Cover Catch Digraphs. <https://cran.r-project.org/package=cccd>. Accessed: 2019-12-18.
- Ministerio de Agricultura, Pesca y Alimentación (2019). Informe del consumo alimentario en España 2018. <https://www.mapa.gob.es/gl/alimentacion/temas/consumo-y-comercializacion-y-distribucion-alimentaria/panel-de-consumo-alimentario/ultimos-datos/> Accessed: 2019-12-02.
- Moelich, E. I., Muller, M., Joubert, E., Næs, T., & Kidd, M. (2017). Validation of projective mapping as potential sensory screening tool for application by the honeybush herbal tea industry. *Food Research International*, 99, 275–286. <https://doi.org/10.1016/j.foodres.2017.05.014>.
- Moussaoui, K. A., & Varela, P. (2010). Exploring consumer product profiling techniques and their linkage to a quantitative descriptive analysis. *Food Quality and Preference*, 21(8), 1088–1099. <https://doi.org/10.1016/j.foodqual.2010.09.005>.
- Næs, T., Berget, I., Liland, K. H., Ares, G., & Varela, P. (2017). Estimating and interpreting more than two consensus components in projective mapping: INDSCAL vs. multiple factor analysis (MFA). *Food Quality and Preference*, 58, 45–60. <https://doi.org/10.1016/j.foodqual.2016.11.012>.
- Nestrud, M. A., & Lawless, H. T. (2008). Perceptual mapping of citrus juices using projective mapping and profiling data from culinary professionals and consumers. *Food Quality and Preference*, 19(4), 431–438. <https://doi.org/10.1016/j.foodqual.2008.01.001>.
- Nestrud, M. A., & Lawless, H. T. (2010). Perceptual mapping of apples and chesnuts using projective mapping and sorting. *Journal of Sensory Studies*, 25, 390–405. <https://doi.org/10.1111/j.1745-459X.2009.00266.x>.
- Orden, D., & Fernández-Fernández, E. (2020). “Data for: Geometric and statistical techniques for projective mapping of chocolate chip cookies with a large number of consumers”, Mendeley Data, v1. <http://dx.doi.org/10.17632/msm2j99rcm.1>.
- Orden, D., & Tejedor-Romero, M. (2019). SensoGraph (under registration process). <https://sensograph.it>. Accessed: 2019-12-17.
- Orden, D., Fernández-Fernández, E., Rodríguez-Nogales, J. M., & Vila-Crespo, J. (2019). Testing SensoGraph, a geometric approach for fast sensory evaluation. *Food Quality and Preference*, 72, 1–9. <https://doi.org/10.1016/j.foodqual.2018.09.005>.
- Page, J. (2005). Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley. *Food Quality and Preference*, 16(7), 642–649. <https://doi.org/10.1016/j.foodqual.2005.01.006>.
- R Development Core Team (2007). R: A language and environment for statistical computing. Available at <http://www.R-project.org>.
- Risvik, E., McEwan, J. A., Colwill, J. S., Rogers, R., & Lyon, D. H. (1994). Projective mapping: A tool for sensory analysis and consumer research. *Food Quality and Preference*, 5(4), 263–269. [https://doi.org/10.1016/0950-3293\(94\)90051-5](https://doi.org/10.1016/0950-3293(94)90051-5).
- Savidan, C. H., & Morris, C. (2015). Panelists’ Performances and Strategies in Paper-Based and Computer-Based Projective Mapping. *Journal of Sensory Studies*, 30(2), 145–155. <https://doi.org/10.1111/joss.12146>.
- Shao, J., & Tu, D. (1995). *The jackknife and bootstrap* New York: Springer-Verlag. <https://doi.org/10.1007/978-1-4612-0795-5>.
- Tarrega, A., Marcano, J., & Fiszman, S. (2017). Consumer perceptions of indulgence: A case study with cookies. *Food Quality and Preference*, 62, 80–89. <https://doi.org/10.1016/j.foodqual.2017.07.001>.
- Teillet, E., Schlich, P., Urbano, C., Cordelle, S., & Guichard, E. (2010). Sensory methodologies and the taste of water. *Food Quality and Preference*, 21(8), 967–976. <https://doi.org/10.1016/j.foodqual.2010.04.012>.
- The Pandas Development Team (2020). pandas-dev/pandas: Pandas <https://doi.org/10.5281/zenodo.3509134>.
- The scikit-bio development team (2020). scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers. <http://scikit-bio.org>.
- Tomic, O., Berget, I., & Næs, T. (2015). A comparison of generalised procrustes analysis and multiple factor analysis for projective mapping data. *Food Quality and Preference*, 43, 34–46. <https://doi.org/10.1016/j.foodqual.2015.02.004>.
- Valentin, D., Cholet, S., Nestrud, M., & Abdi, H. (2016). Projective mapping and sorting tasks. In J. Hort, S. Kemp, & T. Hollowood (Eds.). *Descriptive Analysis in Sensory Evaluation* (pp. 1–19). Wiley-Blackwell.
- Varela, P., & Ares, G. (2012). Sensory profiling, the blurred line between sensory and consumer science. A review of novel methods for product characterization. *Food Research International*, 48(2), 893–908. <https://doi.org/10.1016/j.foodres.2012.06.037>.
- Varela, P., & Ares, G. (2014). *Novel techniques in sensory characterization and consumer profiling*. CRC Press.
- Veinand, B., Godefroy, C., Adam, C., & Delarue, J. (2011). Highlight of important product characteristics for consumers. Comparison of three sensory descriptive methods performed by consumers. *Food Quality and Preference*, 22(5), 474–485. <https://doi.org/10.1016/j.foodqual.2011.02.011>.
- Vidal, L., Cadena, R. S., Antúnez, L., Giménez, A., Varela, P., & Ares, G. (2014). Stability of sample configurations from projective mapping: How many consumers are necessary? *Food Quality and Preference*, 34, 79–87. <https://doi.org/10.1016/j.foodqual.2013.12.006>.
- Vidal, L., Jaeger, S. R., Antúnez, L., Giménez, A., & Ares, G. (2016). Product spaces derived from projective mapping and CATA questions: Influence of replicated assessments and increased number of study participants. *Journal of Sensory Studies*, 31, 373–381. <https://doi.org/10.1111/joss.12220>.
- YouGov (2019). Italian cuisine is world’s most popular. <https://yougov.co.uk/topics/food/articles-reports/2019/03/12/italian-cuisine-worlds-most-popular> Accessed: 2019-12-18.

The free-linking task

” *Freedom. To those that don’t have it, it’s more valuable than gold. But where should it start and end? We humans often think we have the right to expand, absorb, convert, or possess anything we need to reach our dreams.*

— **Adam Jensen**

(Deus Ex: Human Revolution)

Free-sorting is another method for opinion profiling. Opposed to projective mapping, which expects participant to place samples in a two-dimensional space depending on their similarity, sorting is based on a simpler idea. Participants have to divide the set of samples into groups. This idea has two drawbacks: similarity is not binary and does not behave in a completely transitive way.

In terms of graph theory, the second disadvantage of sorting can be expressed with edges. The existence of edges $\{A, B\}$ and $\{B, C\}$ force the existence of the edge $\{A, C\}$ since A, B, C would belong in the same group. However, that edge might not be intended to exist. For this reason, we propose another graph-based alternative called free-linking. It allows participants to freely create or delete edges between the vertices which represent different items in the analysis session. The dataset of graphs is processed using DISTATIS, the same statistical framework used to process free-sorting data, but it could also be processed using *SensoGraph*, extracting partial matrices from free-linking graphs.

In this article, we analyze the similarity of a common set of samples using both free-sorting and free-linking. The conclusion of the study shows that letting participants define edges in the graph freely allows a higher grade of precision in the opinion analysis. This model of similarity is more realistic compared to the transitivity of sorting. Therefore, free-linking is proposed as an interesting alternative to be considered in the future, leading to finer-grained results.



Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

The free-linking task: A graph-inspired method for generating non-disjoint similarity data with food products

Jacob Lahne^{a,*}, Katherine Phetxumphou^a, Marino Tejedor-Romero^b, David Orden^b^a Department of Food Science & Technology, Virginia Tech, 1230 Washington St SW, Blacksburg, VA 24060, USA^b Departamento de Física y Matemáticas, Universidad de Alcalá, Campus Universitario, Ctra. Madrid-Barcelona, Km. 33.600, 28805 Alcalá de Henares, Spain

ARTICLE INFO

Keywords:

Sensometrics

Free sorting

Free linking

DISTATIS

Graph theory

Network theory

ABSTRACT

“Free sorting”, in which subjects are asked to sort a set of items into groups of “most similar” items, is increasingly popular as a technique for profiling sets of foods. However, free sorting implies an unrealistic model of sample similarity: that similarity is purely binary (is/is not similar) and that similarity is fully transitive (similarities {A, B} and {B, C} imply {A, C}). This paper proposes a new method of rapid similarity testing—the “free-linking” task—that solves both problems: in free linking, subjects draw a *similarity graph* in which they connect pairs of samples with a line if they are similar, according to the subject’s individual criteria. This simple task provides a more realistic model of similarity which allows degrees of similarity through the *graph distance* metric and does not require transitive similarity. In two pilot studies with spice blends (10 samples, 58 subjects) and chocolate bars (10 samples, 63 subjects), free linking and free sorting are evaluated and compared using DISTATIS, RVb, and the graph parameters *degree*, *transitivity*, and *connectivity*; subjects also indicated their preferences and ease-of-use for the tasks. In both studies, the first two dimensions of the DISTATIS consensus were highly comparable across tasks; however, free linking provided more discrimination in dimensions three and four. RVb stability was equivalent for the two methods. Graph statistics indicated that free linking had greater discrimination power: on average subjects made similarity groupings with lower degree, lower transitivity, and higher connectivity for free linking in both studies. However, subjects did overall find free sorting easier and liked it more, indicating a higher cognitive difficulty of free linking. The free-linking task, therefore, provides more robust, realistic similarity maps at the cost of higher panelist effort, and should prove a valuable alternative for rapid sensory assessment of product sets.

1. Introduction

Methods for rapidly identifying similarities and differences in sets of food products have become increasingly popular in sensory evaluation (Delarue, 2015; Valentin et al., 2012; Varela & Ares, 2014). In particular, “free sorting”, in which subjects are asked to sort a set of items (in this case, foods or beverages) into groups of “most similar” items is increasingly popular as a technique for profiling sets of foods (e.g., Lahne et al., 2018). Free sorting presents several advantages: it does not require that subjects be trained, it is sensitive and stable with relatively low numbers of subjects (usually as low as 25 subjects), it can accommodate relatively high numbers of samples (as many as 20), and it has been shown to give product “maps” or “configurations” (through multivariate analyses) that bear a close resemblance to those from traditional and more work-intensive methods like Descriptive Analysis.

Furthermore, unlike other rapid methods like Projective Mapping or Flash Profiling (Dehlholm et al., 2012), free sorting only requires that subjects make simple, holistic decisions of similarity or difference, rather than requiring a scaled degree of difference that may induce a higher cognitive load.

However, a key disadvantage of free sorting is that the task of sorting samples makes some strong assumptions about the underlying similarities between the products that are being modeled. Groups in free sorting are *disjoint*, meaning that no element can belong to two groups. Given samples A, B, and C there is no way that the same subject can create two similarity sets such as {A, B} and {B, C} without creating a superset {A, B, C}. This simplifies the task for the subjects and reduces the time and amount of samples required (because retasting is minimized), but this restriction has two potentially undesirable consequences. The first is that the same subject cannot represent different *types* or *dimensions* of

* Corresponding author.

E-mail address: jlahne@vt.edu (J. Lahne).<https://doi.org/10.1016/j.foodqual.2021.104355>

Received 13 April 2021; Received in revised form 4 August 2021; Accepted 5 August 2021

Available online 17 August 2021

0950-3293/© 2021 Elsevier Ltd. All rights reserved.

similarity in the same sort: it is easily conceivable that A and B are similar in terms of one attribute, say, “sweetness”, while A and C are similar in terms of another, say “appearance”. It is quite easy to imagine real-world situations in which this occurs. The second consequence is that similarity is necessarily modeled as fully transitive: if A is similar to B, and B is similar to C, then A must be similar to C, and furthermore the data can only indicate that all three samples *are equally similar*. This is also clearly contrary to easily imagined real circumstances: perhaps A, B, and C are all “sweet”, but while A and B are equally sweet, C is only half as sweet. Should a single subject be required to group these together?

Two closely related alternatives have been suggested for the simple free-sorting task that address these issues: free *multiple*-sorts (Blanchard & Banerji, 2016; Dehlholm, 2015; Dehlholm et al., 2012) and *hierarchical* free-sorts (Koenig et al., 2020, 2021). The former modification asks subjects, after they have completed a simple free-sorting task, to repeat the task until they feel they have exhausted all possible grouping configurations (Dehlholm, 2015); the latter asks subjects, once they completed a simple free-sorting task, to continue making groups of *groups* until they cannot proceed further (Koenig et al., 2021). Thus, free multiple-sorting solves the first problem highlighted above, and hierarchical free-sorting solves the second problem. However, neither approach solves *both* problems, and they both introduce problems of panelist motivation, in that they require a much more extensive data-collection procedure that will be discouraging for some subjects. This is a more major problem when a large number of samples is used, as in Koenig et al. (2020), but difficulty and motivation problems are reported with as few as 18 complex samples sorted by taste (Kessinger et al., 2020). In addition, the data collection for both methods is much more complicated and more poorly supported in practical data-management programs (based on the authors’ personal communications with major sensory and survey software providers in pursuit of these methods), which appears to have limited the adoption of either approach in academia and industry in favor of the simple free-sorting task. For example, Spencer et al. (2016) had to write custom software to support hierarchical free-sorting, and authors as recent as Koenig et al. (2020, 2021) have used paper ballots because of the lack of software supporting hierarchical free-sorting, requiring extensive transcription of results.

Therefore, in this manuscript we propose an alternative task to the free-sorting task, inspired by graph theory (Gross et al., 2014), which we term the “free-linking” task. In the free-linking task, subjects are given a set of samples just as in free sorting, but rather than forming disjoint groups, subjects are asked to indicate, for each pair of samples, whether the samples are similar. This connect-the-dots interface was implemented in the SensoGraph system (Orden et al., 2019, Alcalá, ES) in order to support this task, in which subjects are asked to draw “links” between samples if they are similar (Fig. 1). However, a paper-based system for free linking would be no harder to implement than a paper-based simple free-sorting task.

While the free-linking task solicits binary similarity data on a pairwise basis for samples—two samples are either similar or they are not—it does not impose the disjoint, restrictive model of similarity implied by free sorting. Given 3 samples A, B, and C it is possible for a subject to indicate, pairwise, that there are similar pairs {A, B} and {B, C} without indicating that A and C are directly similar. Put another way, the free-linking task asks each subject to draw their own similarity graph for the samples (Lahne, 2020; Orden et al., 2019; 2021). Unlike previous graph-based approaches to similarity in food products, where just the presence or absence of a connection was considered, in free linking we make use of the *graph distance* between samples as a basis for a dissimilarity matrix for further analysis (Chartrand & Zhang, 2014). In the example above, $distance(A, B) = distance(B, C) = 1$, while $distance(A, C) = 2$. This allows the analyst to *infer* from a single subject’s data that, for the example above, there might be some shared similarity between A and C without the link {A, C} actually being drawn. This same change also addresses the second problem with simple free-sorting: subjects can now indicate pairwise whether samples are similar, but because there

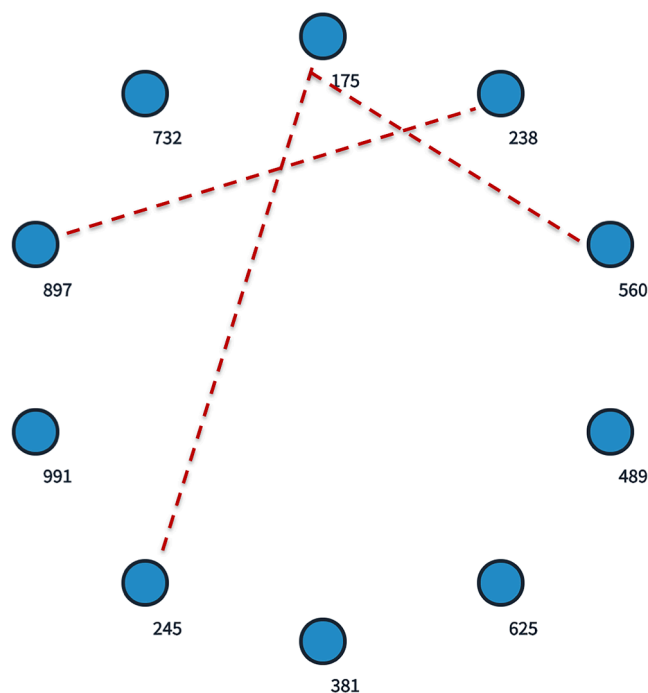


Fig. 1. Interface for individual subjects’ free-linking task, as rendered in SensoGraph (Orden et al., 2019). Note that sample order is randomized between subjects.

are not larger similarity groups (e.g., {A, B, C} in free sorting) it is not required that all samples that are connected be similar *in the same way*. This allows more flexibility for a subject’s holistic similarity judgments (Fig. 2; see also Fig. 3 for details on the dissimilarity).

The free-linking task can be analyzed by the same tools that exist for the free-sorting task: dimensionality reduction (through MDS, DISTATIS, and other approaches) and graph-based approaches like Sorting Backbone Analysis. This allows analysts used to free sorting to easily employ free linking, and for direct comparison of results.

Therefore, it is reasonable to hope that the free-linking task will provide results that are comparable to free-sorting in terms of ease of deployment and data collection, but might allow for more realistic and detailed results. In particular, the lack of forced memberships to a group should allow for easier distinction among similar but not identical samples—that is, a more multidimensional structure of similarity and difference. In order to investigate the utility of the free-linking task, we report the results of two pilot studies in which subjects used both free sorting and free linking to report their perceptions of different food products. In both pilot studies subjects completed both free-sorting and free-linking tasks for the same samples in a counterbalanced order. In the first study, subjects evaluated 10 blends of 4 dried spices (cinnamon, turmeric, pepper, and cardamom) for similarity by aroma. In the second study, subjects evaluated 10 commercial chocolate samples for similarity by taste. We hypothesized that the overall similarity configuration should be similar between the two methods, and that the results of the two methods should be equally stable, but that the free-linking results would provide more realistic, multidimensional models of similarity, which should be evident in parameters for the graphs derived from the similarity measurements as well as in visualizations from DISTATIS.

2. Materials and methods

The two studies reported were very similar in most details besides sample type, and so the basic information distinguishing the studies is given below, followed by details on methodology and analysis that were the same for both studies.

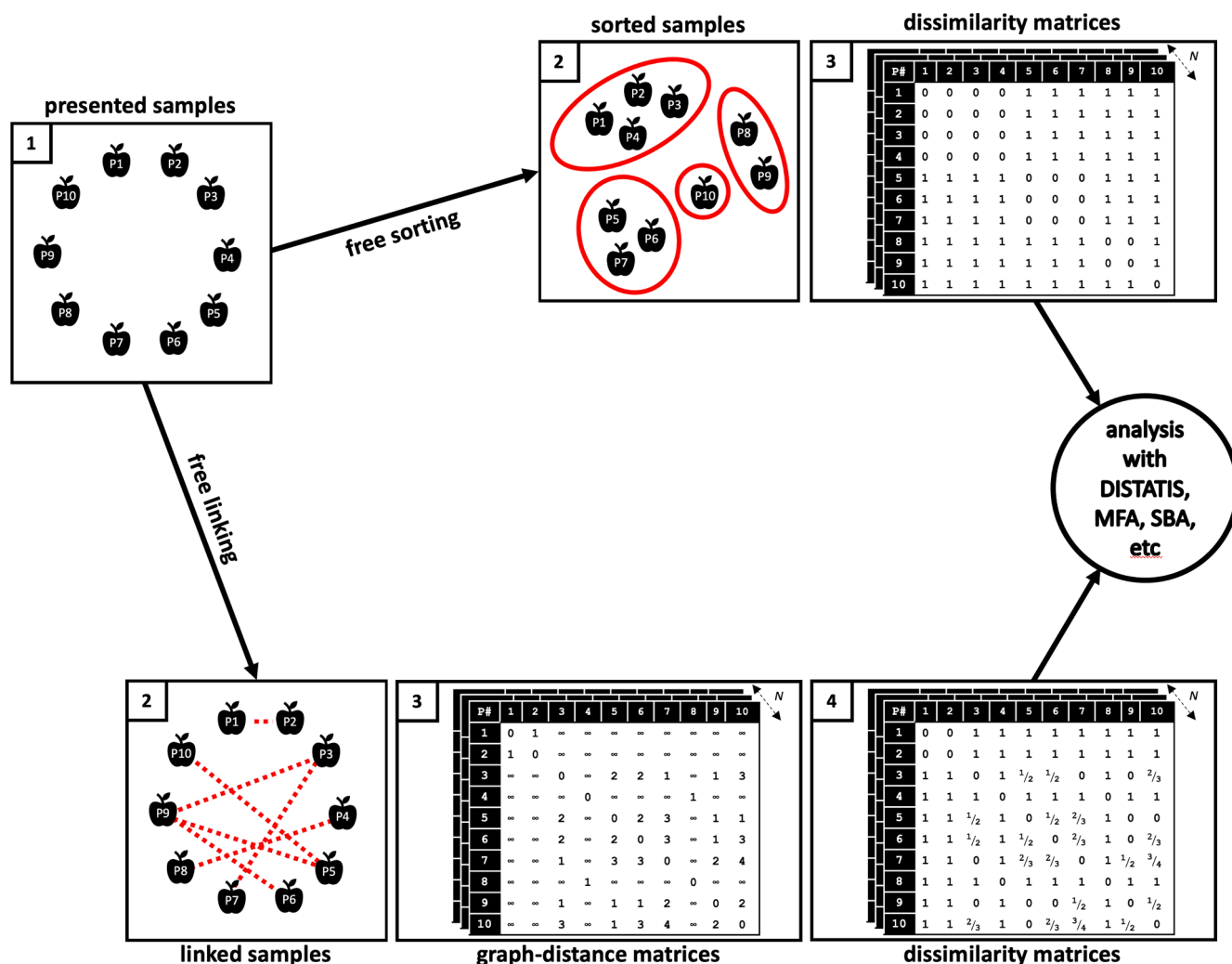


Fig. 2. Schematic representation of free sorting (top) and free linking (bottom). From the same samples (presented in random order to each subject) in (1), the methods diverge. For free sorting, subjects group samples (2) and their groupings are transformed directly to binary dissimilarities (3). For free linking, subjects indicate pairwise similarity (2), which is transformed into graph distances (3), and then to [0,1]-range dissimilarity (4, with details given in Fig. 3). At this point, the same analyses can be conducted on the each of the dissimilarity matrices.

2.1. Study 1–Spice sorting

Study 1 was conducted in November and December of 2019, and used spices and spice blends as stimuli. Sample details are given in Table 1. All spices were purchased at Kroger (Blacksburg, VA, see Table 1). Samples were presented to subjects in foil-wrapped glass vials in order to avoid visual discrimination, and evaluation was entirely orthonasal.

A total of $N = 58$ subjects (38 female, 20 male, average age 29 years old) participated in Study 1. Subjects were recruited from the Virginia Tech/Blacksburg community. Subjects were not trained sensory panelists (e.g., for Descriptive Analysis), but some had participated in previous untrained sensory tests at Virginia Tech. Subjects received no compensation, but were given snacks after completing Study 1.

2.2. Study 2–Chocolate sorting

Study 2 was conducted in November of 2020, and used commercial chocolate bars as stimuli. Sample details are given in Table 1. All chocolate bars were purchased at Kroger (Blacksburg, VA, see Table 1). Samples were presented in souffle cups with the bars' identifying details (e.g., logos) effaced, in natural light, and evaluation was by taste and

retronasal flavor.

A total of $N = 63$ subjects (49 female, 14 male, average age 34 years old) participated in Study 2. Subjects were recruited from the Virginia Tech/Blacksburg community. Subjects were not trained sensory panelists (e.g., for Descriptive Analysis), but some had participated in previous untrained sensory tests at Virginia Tech, including some who had participated in Study 1. Subjects received no compensation, but were given snacks after completing Study 2.

2.3. Overall study design

Both studies used the same overall design. Subjects were recruited to participate in free-linking and free-sorting of the same samples. In order to obtain within-subjects data, subjects were randomly assigned one of the two tasks first, then took a short break, then completed the other of the two tasks, then completed a short survey that asked them about their perceptions of the tasks and some basic demographic details. In both free-sorting and free-linking studies, subjects were seated at tables with a 36" x 36" workspace available and allowed to organize their samples spatially prior to entering their judgments into the data-collection software.

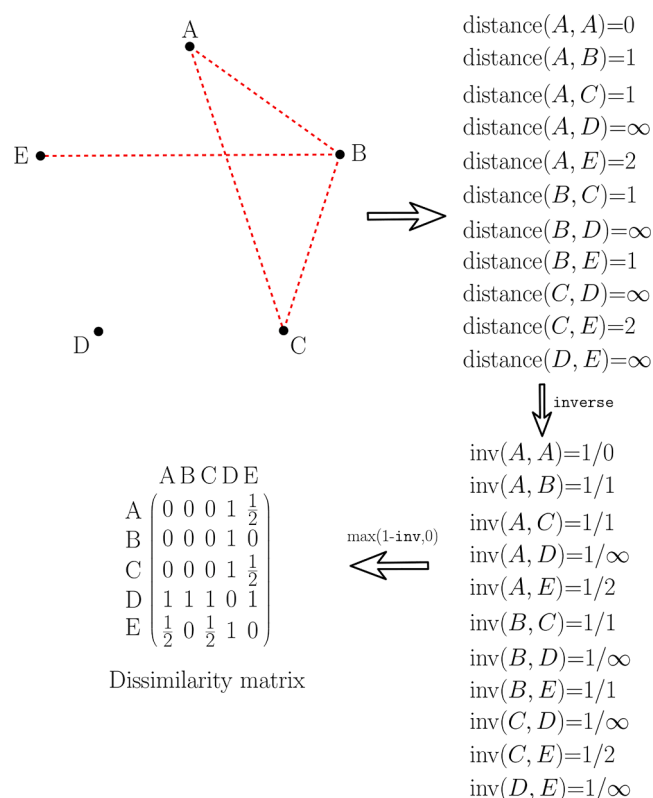


Fig. 3. Schematic for deriving dissimilarity from graph distance, based on Koenig et al. (2021).

Table 1

Sample information for Study 1 and Study 2.

Study 1 – Spices*		
Sample Name	Recipe	
Cinnamon	1 g ground cinnamon	
Cardamom	1 g ground cardamom	
Pepper	1 g ground black pepper	
Turmeric	1 g ground turmeric	
Cinnamon + cardamom	0.5 g ground cinnamon + 0.5 g ground cardamom	
Cinnamon + pepper	0.5 g ground cinnamon + 0.5 g ground black pepper	
Cinnamon + turmeric	0.5 g ground cinnamon + 0.5 g ground turmeric	
Cardamom + pepper	0.5 g ground cardamom + 0.5 g ground black pepper	
Cardamom + turmeric	0.5 g ground cardamom + 0.5 g ground turmeric	
Pepper + turmeric	0.5 g ground black pepper + 0.5 g ground turmeric	
Study 2 - Chocolate		
Manufacturer	Chocolate type	Cocoa content
Cadbury	Dark	35%?
Hershey's	Dark	45%?
Green & Black's	Dark	70%
Endangered Species	Dark	72%
Green & Black's	Dark	85%
Pascha	Dark	85%
Cadbury	Milk	26%?
Hershey's	Milk	30%?
Green & Black's	Milk	34%
Endangered Species	Milk	48%

* All spices are McCormick Gourmet Organic line ground spices (no whole spices were used for the purpose of blending the recipes).

[?] Information gathered indirectly from manufacturer's website rather than packaging.

2.4. Free-sorting task

In the free-sorting task, subjects received all 10 samples at the same time in a randomized order. Sorting data was collected using the

Compusense Cloud (Guelph, ON) system. Subjects were prompted to “sort into groups based on similarities”. They were informed that there was no right answer, and told that they could make any number of groups between two (2) and nine (9), with as many samples as they chose in each group.

2.5. Free-linking task

In the free-linking task, subjects received all 10 samples at the same time in a randomized order, positioned as the vertices of a regular polygon (see Figs. 1 and 2). Linking data was collected using the SensoGraph (Orden et al., 2019, Alcalá, ES) system. Subjects were prompted to “join with a line those pairs of products you consider similar, dragging from one to the other with the finger or the mouse” (see Fig. 1). The codes presented on the screen for the SensoGraph interface were given in random order for each subject. Subjects were able to remove lines they had previously made (in case of mistakes or revisions in judgment) before submitting their answers.

2.6. Data Analysis

Results from both free sorting and free linking were analyzed in parallel in order to compare the results of the method. This parallelism is enabled by the data structure provided by both methods: the dataset for each analysis is an $N \times K \times K$ array of (dis)similarity matrices, where N is the number of subjects and K is the number of samples. In free sorting, each $K \times K$ slice is composed by cell entries a_{ij} which are binary (either 0 or 1), representing whether, for the current subject, samples i, j were sorted together. The raw data is a *similarity* measure in which a 1 indicates similarity through group membership, and the dissimilarity matrix, which is obtained by subtracting every entry from 1, can be treated as binary distance and is analyzed via MDS or DISTATIS (Abdi et al., 2007). In free linking, the graph drawn by the current subject provides a graph distance between each pair of samples i and j , as an integer between 1 (if the connection $\{i, j\}$ is present) and ∞ (if there is no path between i and j on the graph). The raw graph distance is the number of edges comprising the shortest path between the two pairs of samples in the graph (see Figs. 2 and 3). For the dissimilarity matrix actually analyzed by DISTATIS we adapt the cophenetic dissimilarity from Koenig et al. (2021, see Figs. 2 and 3): the corresponding cell entry a_{ij} of the $K \times K$ slice is defined as the subtraction from 1 of the inverse of the graph distance between i and j (defining $1/\infty$ as 0, and setting a minimum of 0 for dissimilarity of a sample with itself or with samples to which it is directly linked), so that the cell entries a_{ij} are no longer binary but range in the interval $[0, 1]$, with larger values indicating lower similarity and smaller values standing for higher similarity. The diagonal of the matrix is set to 0, indicating that all samples are identical with themselves as would be expected for a distance matrix.

Data were first analyzed by DISTATIS in order to compare consensus similarity configurations for samples across methods (Abdi et al., 2007). Confidence ellipses were generated through bootstrapping (Beaton et al., 2013). A key property of any rapid sensory method is how well samples and groups of samples are distinguished: this is clearly related to (but also not identical to) discrimination ability for the method. Examination of product separation on the first four DISTATIS axes for both methods via actual observations as well as bootstrapped confidence intervals were considered as evidence. Choice of 4 axes for examination (out of a possible 10 for each sample) were motivated by examination of scree plots for the DISTATIS S_+ matrices (Abdi et al., 2007, not shown) as well as by general practice in industry and the literature for “significant dimensions” for interpretation.

The stability of results for a given number of subjects—that is, the required number of subjects—for each method was evaluated through a bootstrapping approach to simulate panels of different sizes and compare these simulated results to the actual, observed results. Specif-

ically, generalized stability, termed RVb by [Blancher et al. \(2012\)](#), was calculated for free sorting and free linking: bootstrapped samples of subjects, of sizes 2 to N (where N is the number of subjects in the particular study) were drawn (with $i = 100$ replicates at each sample size), and the average RV between the DISTATIS F (factor score) matrices from the bootstrap sample and the full dataset was calculated at each sample size. [Blancher et al. \(2012\)](#) recommend that stability can be considered achieved at the number of subjects for which the bootstrapped average RVb exceeds 0.95.

Graph theory was also used to evaluate whether individual subjects' free-sorting and free-linking groupings were in fact different. For

sorting, each individual's $K \times K$ slice was treated as the (symmetric) adjacency-matrix representation of an undirected graph ([Gross et al., 2014](#)). For linking, the undirected graph drawn by each individual was used. In each subject's graph, the nodes represent the samples, and an edge between two nodes indicates that the subject sorted or linked two samples as similar ([Lahne, 2020](#); [Orden et al., 2019](#)). This graph representation provides several simple parameters that give insight into the similarity structure.

The degree of each node indicates how many edges are incident to it ([Gross et al., 2014](#)); thus, in sorting or linking higher degree for a node means the corresponding sample was considered similar to more other

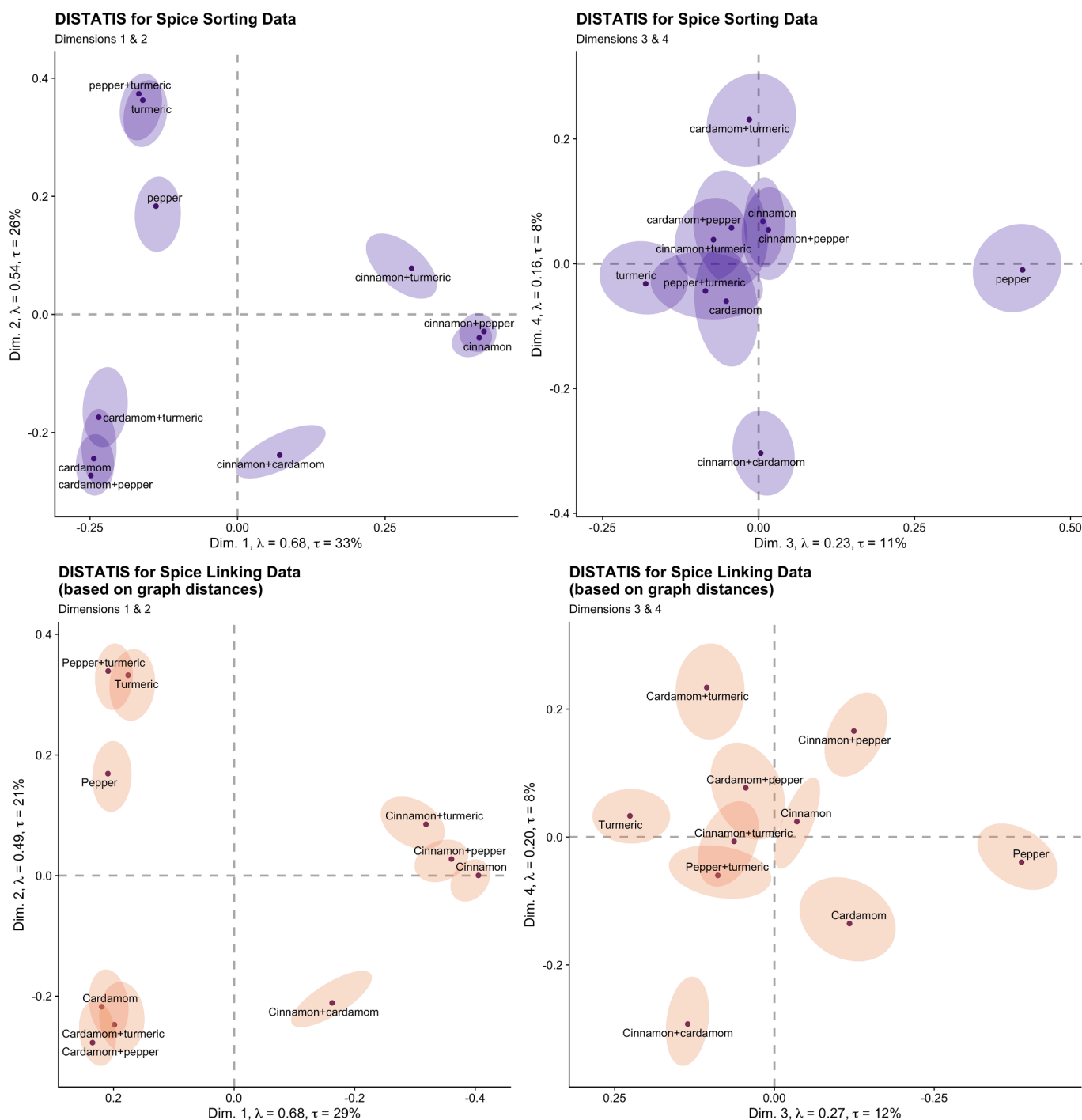


Fig. 4. DISTATIS biplots for free sorting (top, in purple) and free linking (bottom, in orange) of spice-study results. The left-hand column gives Dimensions 1 and 2, while the right-hand column gives Dimensions 3 and 4 of the respective spaces. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

samples. Comparison of average degree per subject and sample for each method gives an indication of discrimination capacity: higher average degree indicates less discrimination between samples, as subjects consider more samples similar.

The *transitivity on triads* (Arney & Horton, 2014) is the fraction indicating, for the total number of node triads A, B, C with connections {A, B}, {B, C}, how many of them also contain the connection {A, C}. In the literature this is also called the graph “clustering coefficient” (Kolaczyk & Csárdi, 2014). In terms of the sorting and linking tasks, this is a measure of the likelihood that similarities {A,B} and {B,C} imply that similarity {A,C} also exists; when transitivity is higher it may

indicate a lower discrimination capability.

The *average connectivity* of a graph (Beineke et al., 2002) is a parameter that measures, in each subject’s results, the average over all pairs of nodes A and B, how many independent paths connect A and B. In the context of sorting and linking, lower average connectivity will be associated with more disjoint groups, which is an indicator of less robust or realistic models of similarity.

Subjects’ preferences for method were evaluated for each study using simple contingency-table measures, and their opinions of the sorting and linking tasks’ ease of use and enjoyability were evaluated using repeated measures ANOVA.

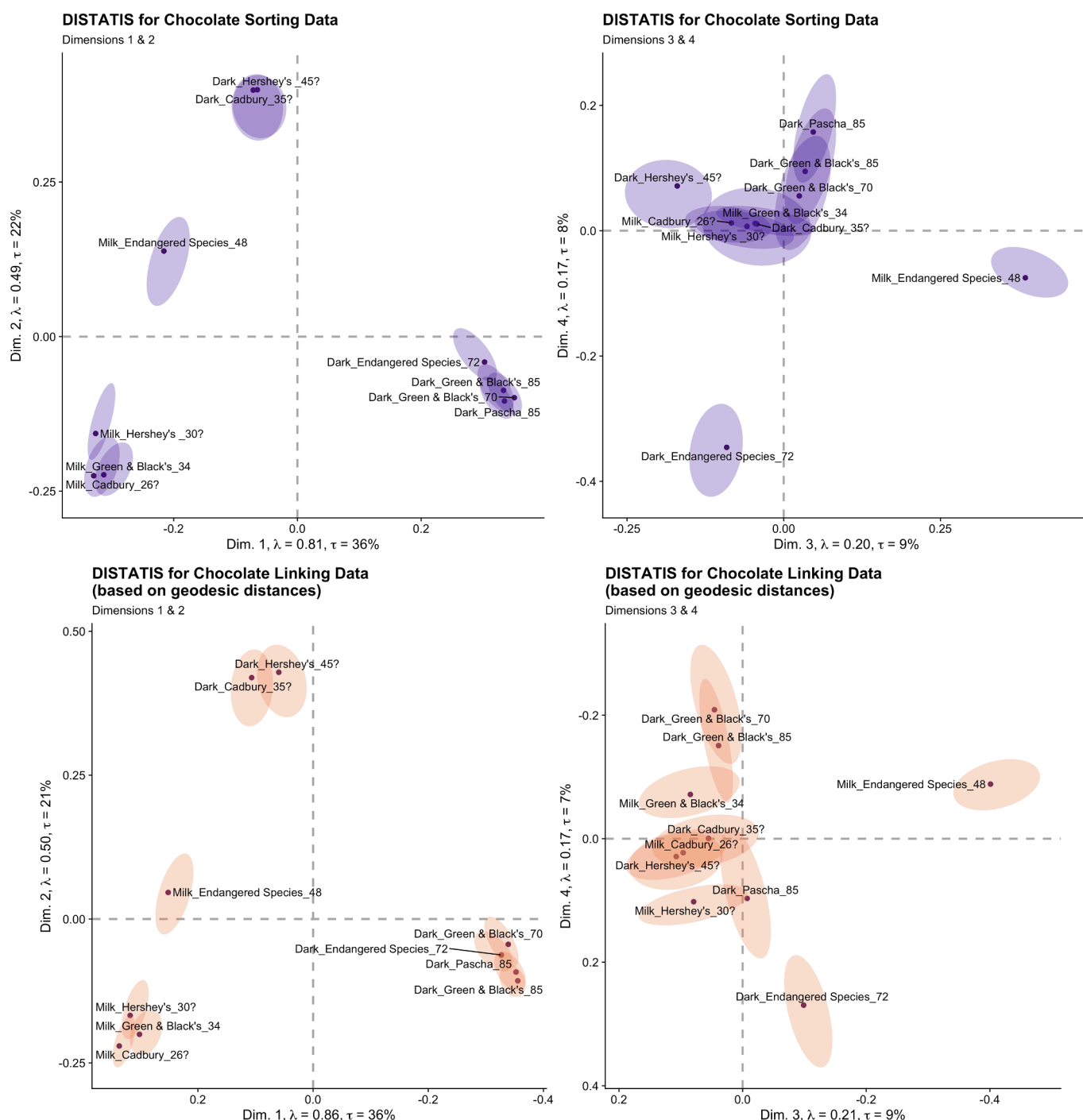


Fig. 5. DISTATIS biplots for free sorting (top, in purple) and free linking (bottom, in orange) of chocolate-study results. The left-hand column gives Dimensions 1 and 2, while the right-hand column gives Dimensions 3 and 4 of the respective spaces. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Data analyses were conducted in R (version 4.0.2). Code for analyses is available from the corresponding author upon request.

2.7. Ethics statement

All research methods were reviewed and approved by the Virginia Tech Human Research Protection Program (IRB # 19-1030).

3. Results

3.1. Product configurations (via DISTATIS)

The overall DISTATIS results for both the spice samples (Study 1) and the chocolate samples (Study 2) are quite similar (Figs. 4 and 5). In the first 2 dimensions of the DISTATIS solutions the configurations of samples are almost identical, although it is worth noting that the derived distances among samples in the chocolate study are larger (Fig. 5). However, for both studies it is apparent that the 3rd and 4th dimensions of the solution contain more valuable discrimination information for free linking than for free sorting. In each case, more samples are clearly discriminated (as can be seen from non-overlapping confidence ellipses) by subjects using free linking than by subjects using free sorting.

The same basic product differences are identified by both methods, but with better resolution through free linking. For the spices, the first dimension separates cinnamon-containing mixes from the rest of the samples, while the second dimension separates cardamom-containing mixes (in both analyses the cinnamon + cardamom mixture falls in between these groups, with a stronger attraction to the cardamom region on the second axis). The third dimension for both studies separates pepper from the remaining samples, but with free linking it is also possible to infer that pepper is being directly opposed to turmeric-containing samples (Fig. 4). In the fourth dimension, two samples that both contain turmeric are opposed: cardamom + turmeric and

cinnamon + cardamom, but again in the free-linking study several other samples (cinnamon + pepper, cardamom) separate clearly on this dimension).

For the chocolate, the first dimension distinctly separates premium, dark chocolates from milk chocolates, while the second axis separates mass-market dark chocolates (Hershey's and Cadbury's) from the other samples. In the third dimension, the sole premium, milk chocolate (Endangered Species) is separated from the remaining samples, but only in the free-linking study is it clear that this dimension is capturing similarities between both chocolates from this producer (Fig. 5). Finally, the fourth dimension separates the dark chocolate from Endangered Species from the remaining chocolates, but, again, in the free-linking study it is clear that there is more separation on this axis, with a strong separation between the two dark chocolates from Green & Black on this axis as well as separation among the other samples.

3.2. Stability (via RVb)

In order to investigate stability of the solutions as a function of the number of panelists, *RVb* was calculated as described in Blancher et al. (2012). Fig. 6 shows the *RVb* results for free sorting and free linking. As is apparent, the desired level of stability (the 0.95 level) is achieved with essentially the same number of subjects for both sorting and linking—although an average of about 1 subject less is required for stability in free sorting than in free linking. Given that this level of stability is achieved at between 8 and 10 subjects in these studies, this difference of a single subject is unlikely to be important in practical applications. In contrast to Blancher et al. (2012), we used all 10 dimensions to calculate *RVb*, but results for only Dimensions 1 and 2 (as calculated in Blancher et al. 2012) were almost identical to the full factor bootstraps (results not shown).

This is a quite low number of subjects when compared to those calculated by Blancher et al. (2012)—it corresponds most closely to the

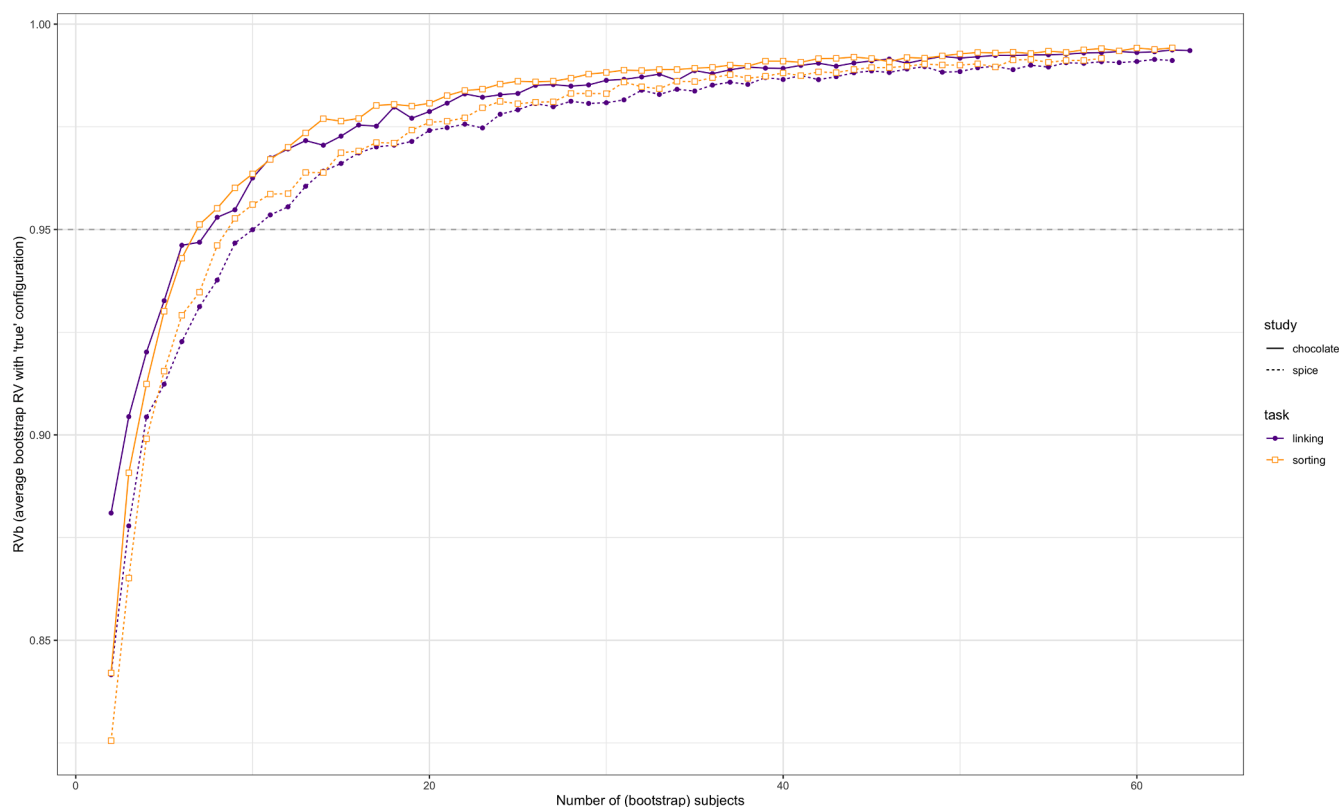


Fig. 6. Stability of consensus solutions as assessed by *RVb* for free linking (purple) and free sorting (orange) in spice (dashed) and chocolate (solid) studies. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

results in that study for a similar dataset of chocolate aromas (DS1, a free sort of 11 samples). While Blancher et al. (2012) do not give details on sample-inclusion criteria, in the case of both Study 1 and Study 2 samples were chosen specifically for their potential to be grouped by subjects (i.e., blends of the same spices and chocolates from the same manufacturers, see Table 1), which may explain the high stability observed here. It is also noticeable that the number of subjects required is slightly lower in Study 2 (chocolate, solid line) than in Study 1 (spice, dashed line). This difference seems like it may be attributed to the difference in modality—taste and flavor for Study 2, and only aroma for Study 1; differences in the products themselves may also be in play. This difference is also evident in the relative size and overlap of confidence ellipses for DISTATIS results (in which the RV coefficient is a key statistic) seen in Figs. 4 and 5. However, there is no evident difference in the RVb patterns between sample type, modality, and methodology (sorting vs. linking). The apparent stability of each method is equivalent.

3.3. Graph parameters

Three key graph parameters were investigated for this study. In a graph, the degree of a node represents the number of incident edges; for the sorting and linking studies, for each subject the degree of each sample indicates the number of other samples to which it was judged similar. Higher degree thus indicates a potentially lower discrimination ability among subjects, as fewer distinctions are made. For both Study 1 and Study 2, the degree distribution for free linking is clearly skewed more right than the degree distribution for free sorting (see Fig. 7). Wilcoxon rank-sum tests indicate that the free-sorting task produces significantly larger degrees per node than the free-linking task for both the spice ($W = 143331, p < 0.05$) and the chocolate ($W = 167328, p < 0.05$) studies. This indicates that free linking better discriminates the samples than free linking.

The *transitivity* on triads of a graph indicates the likelihood, given three nodes A, B, and C and edges {A, B} and {B, C}, that there will also be an edge {A, C}. In terms of free sorting and free linking, transitivity gives another indication of discrimination ability—it is a direct measurement of the degree to which similarities among samples are forced by the method or are allowed to be indicated by the subjects, and ranges

from 0 to 1. In Fig. 8, transitivity is plotted on the Y-axis against degree (see above) on the X-axis. By the nature of the sorting task, transitivity is always 0 or 1; it is only 0 in the degenerate case, when subjects made only pairs of samples, which happened several times in the spice study. For free sorting there is a much broader range of transitivity values in the [0, 1] range, indicating a higher likelihood of actual discrimination by the subjects.

Finally, the *connectivity* of a graph is a measure, for each subject, of the number of distinct, connected paths between all pairs of nodes. Higher connectivity indicates a less disjoint (or disconnected) graph; in terms of free sorting and free linking, lower connectivity would mean more disjoint graphs, which are likely the result of a less realistic similarity model. In Fig. 9, connectivity is plotted on the Y-axis against degree on the X-axis. For both studies, free linking tended to exhibit higher connectivity values than free sorting, as expected, but the differences were in general rather smaller than the differences in connectivity or degree. Thus, while subjects did produce more connected graphs using free linking than free sorting, they did not always produce fully connected graphs.

3.4. Subject preferences

Finally, it is important to consider subjects' experience of the two tasks. In a simple question of overall preference ("Did you prefer the free-sorting or free-linking task?"), panelists preferred free-sorting to free-linking narrowly but insignificantly in Study 1 ($\chi^2_1 = 1.10, ns$), and by a broad and significant margin in Study 2 ($\chi^2_1 = 19.44, p < 0.05$; see Table 2). In neither task did it matter which task the subjects completed first (Study 1: $\chi^2_1 = 0.69, ns$; Study 2: $\chi^2_1 = 1.49, ns$). This can potentially be explained by the difference in complexity of the relative tasks: in Study 1, the test was by aroma only, whereas in Study 2 the subjects had to taste the chocolate. Therefore, it is possible that Study 2 involved a more fatiguing sensory task and a more taxing memory task, and in these circumstances it would make sense that subjects would prefer the simpler free-sorting task, which involves fewer pairwise comparisons. Alternatively, it is possible that the difference may be that the set of samples evaluated in Study 1 was "designed" by blending spices, providing an "easier" similarity structure.

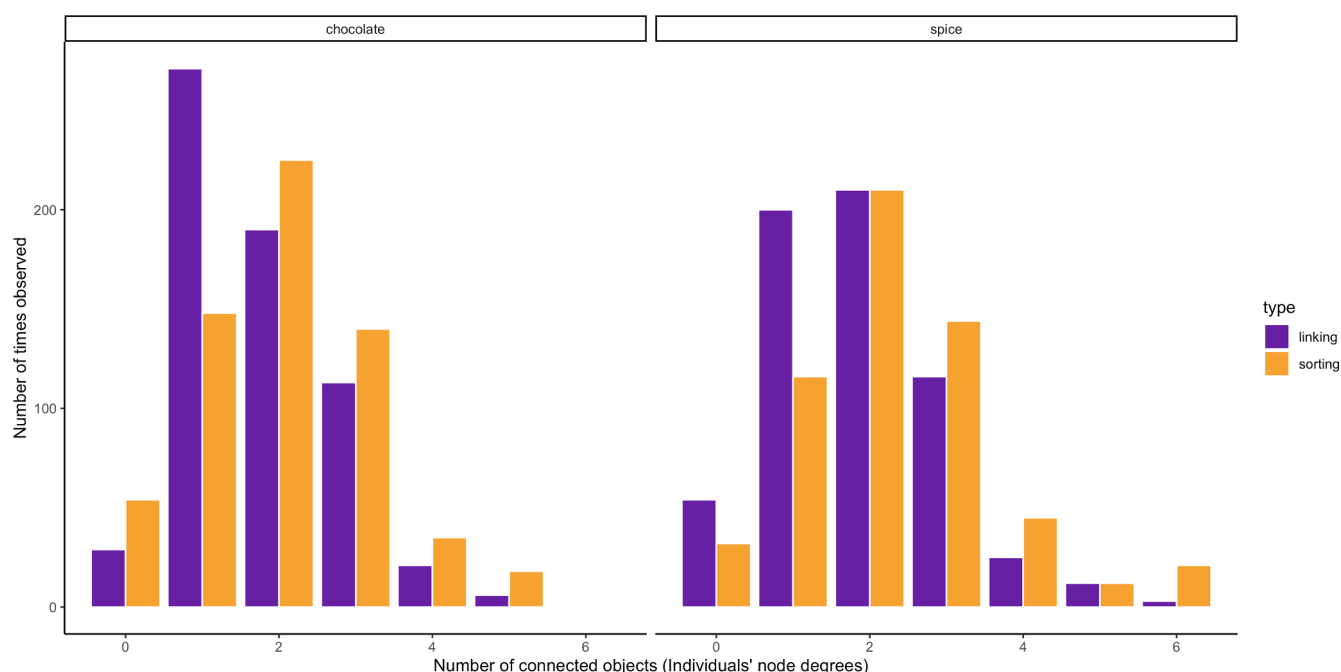


Fig. 7. Degree distributions for spice (left) and chocolate (right) studies for free linking (purple) and free sorting (orange). In these studies, higher degree indicates less power to discriminate among samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

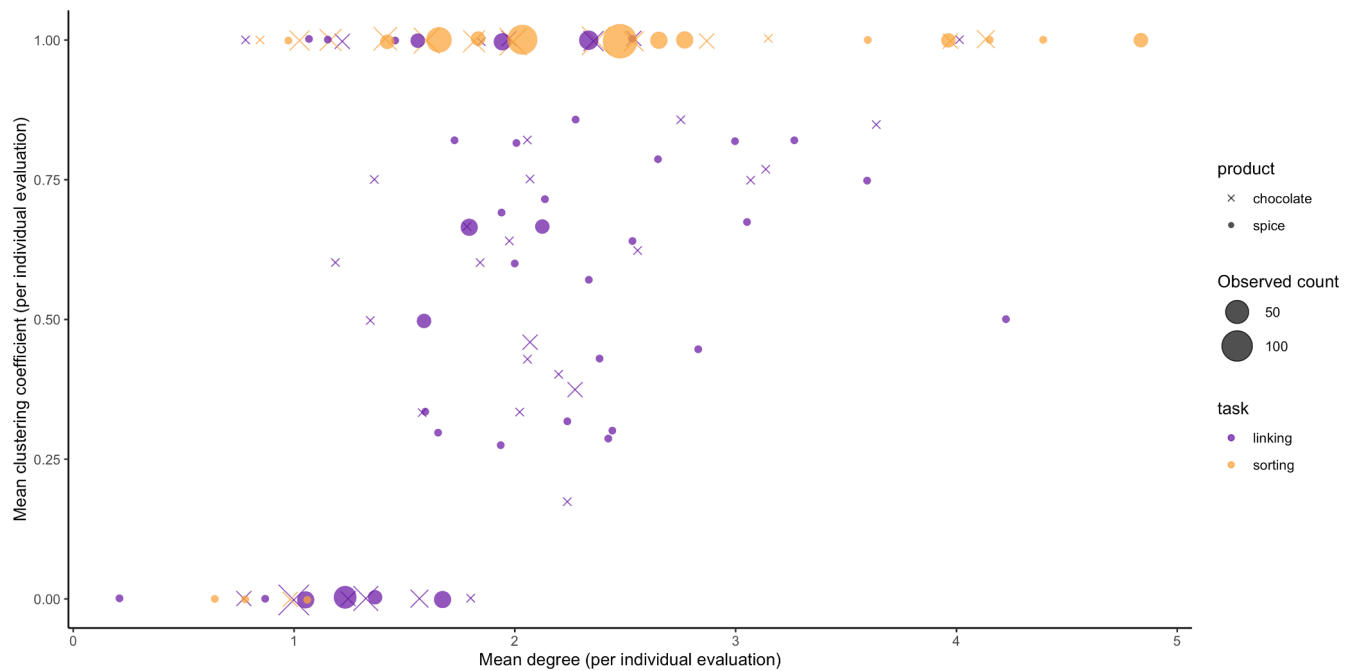


Fig. 8. Scatter plots of individual subjects' degree (with lower degree indicating higher discrimination power) against transitivity (clustering coefficient, with higher values indicating forced grouping/similarity) for free linking (purple) and free sorting (orange). Note that for free sorting, transitivity is *always* equal to 1 except in the rare degenerate case in which subjects only make groups of 2 or fewer samples (bottom left). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

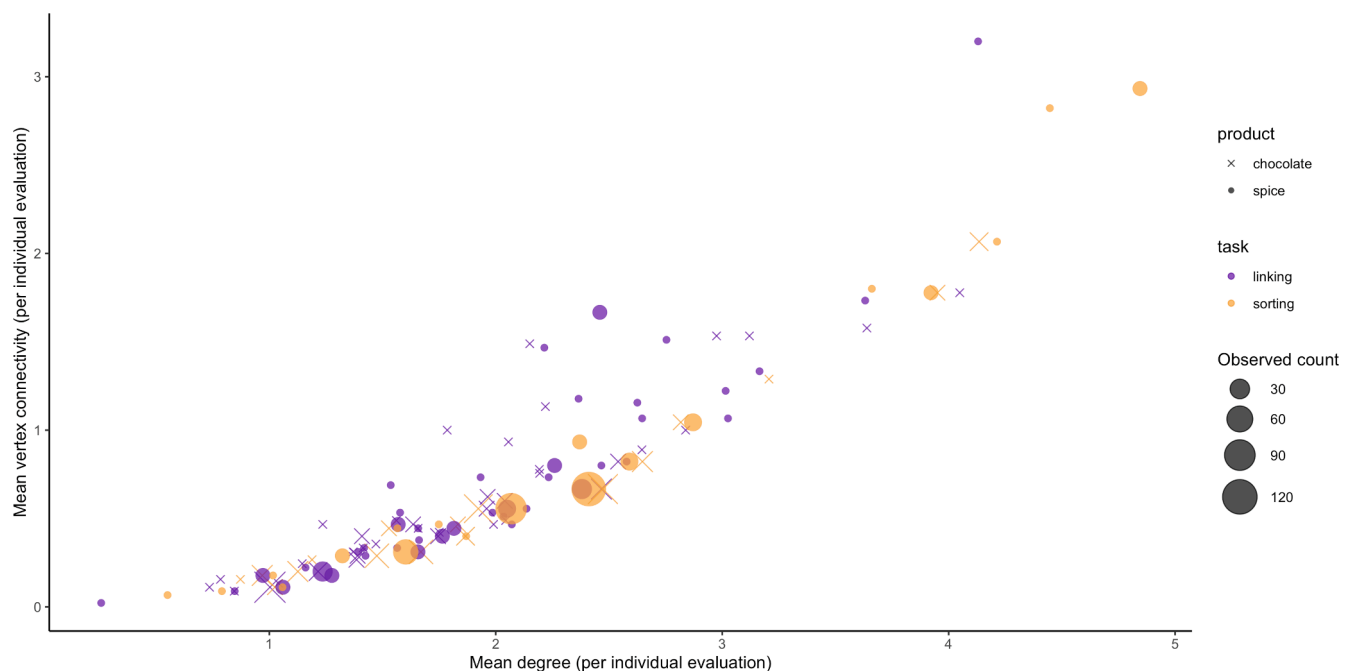


Fig. 9. Scatter plots of individual subjects' degree (with lower degree indicating higher discrimination power) against connectivity (with higher values indicating ability to detect multiple levels of similarity). Note that for free sorting, only high values of degree guarantee higher connectivity, whereas in free linking higher connectivity is achieved at lower degree (with higher discrimination power).

Subjects also answered questions about ease-of-use and rated liking for each task, both on unstructured line scales converted to 10-pt values. Results were analyzed by mixed-effects ANOVA, with the dependent variable (liking or ease) modeled as dependent on the random effect of the particular subject, with the task (free sorting or free linking) as a within-subjects variable and the order of task completion as a between-subjects variable. For all tests, there was no effect of order of task, and no

interaction between order and the task itself, so these results will not be reported in detail. For Study 1, subjects indicated that they did not find any difference in ease-of-use for the two tasks (effect of task on ease-of-use: $F_{1,56} = 0.016, ns$; free sorting $M = 7.62, SD = 1.78$, free linking $M = 7.14, SD = 2.08$), but they did report a significantly higher liking for the free-sorting task (effect of task on liking: $F_{1,56} = 5.14, p < 0.05$; free sorting $M = 7.57, SD = 1.70$, free linking $M = 6.85, SD = 2.08$). For

Table 2

Counts of preference for free-sorting or free-linking task for each study, counted by which test was completed first.

Task Completed First	Prefer Free-Sorting	Prefer Free-Linking
<i>Study 1: Spices (by smell)</i>		
Free-Sorting	15	15
Free-Linking	10	18
<i>Study 2: Chocolate (by taste)</i>		
Free-Sorting	24	10
Free-Linking	25	4

Study 2, subjects indicated significant differences in both ease-of-use (effect of task on ease-of-use: $F_{1,61} = 24.76, p < 0.05$; free sorting $M = 8.36, SD = 1.48$, free linking $M = 6.96, SD = 2.38$) and liking (effect of task on liking: $F_{1,61} = 19.40, p < 0.05$; free sorting $M = 7.54, SD = 1.61$, free linking $M = 6.11, SD = 1.92$). These results can be explained in the same way as the preference results: possibly a significantly higher memory and sensory-fatigue loads for tasting would make free linking a more difficult and less pleasant task than free sorting, or possibly the set of samples evaluated in Study 1 was slightly “easier” than the chocolates in Study 2. In both cases, it is also possible that subjects are simply more familiar with free sorting than with free linking, and familiarity has bred comfort with and preference for that method: while subjects were not surveyed about previous experience, our lab frequently conducts free-sorting studies and some subjects were definitely previous participants.

4. Discussion

Free sorting, as a rapid method for assessing similarities among a set of samples, has become an extremely popular method in both industry and academia (Dehlholm, 2015; Koenig et al., 2020, 2021; Valentin et al., 2012). However, the basic instruction of free sorting—that subjects form disjoint groups according to similarity—implies a model of similarity among the products that is likely to be unrealistic. Specifically, sorting requires that similarities be fully transitive and essentially unidimensional. In contrast, the method of pairwise free-linking, which we have formalized and demonstrated in this paper, provides results that are comparable to free sorting, while avoiding these restrictive assumptions.

In particular, on the same product sets, free linking results in significantly lower vertex degree measurements for each product, indicating that subjects are making more discriminating similarity judgments (Fig. 7). In addition, the transitivity (or “clustering coefficient” Kolaczyk & Csárdi, 2014) of the similarity graphs from free linking were significantly more diverse than those from sorting, which are in general fully transitive (Fig. 8); this explicitly indicates that subjects in free-linking studies are not forced to “close the triangle” when they want to indicate that A and B are similar, as are B and C. At the same time, the connectivity of the free-linking graphs was also noticeably higher than that of the free-sorting graphs (Fig. 9), indicating that individual models of similarity generated through free linking were more robust, with graph distance giving a non-binary similarity measure (Chartrand & Zhang, 2014), which should capture a more multidimensional model of similarity.

This more “multidimensional” similarity is evident in DISTATIS biplots of results of free sorting and free linking on the same samples. Although for both spices (Fig. 4) and chocolate (Fig. 5) gross similarities, represented by Dimensions 1 and 2 of the biplots, are almost identical, there is much better discrimination of samples in Dimensions 3 and 4 for both sample sets. This follows naturally from the two different models of similarity implied by free sorting and free linking. Free sorting emphasizes rapidly finding gross similarities; free linking, while more intensive because of the need for multiple pairwise judgments (Fig. 1), focuses on multidimensional similarity. Nevertheless, both methods provide stable results, as indicated by RVb , at approximately similar numbers of

subjects (Fig. 6). However, it is important to note that, on the whole, subjects found free sorting less taxing and more pleasant than free linking. It will be important to take subject fatigue into account when designing future studies that employ free linking. We might imagine that free linking would also be less fatiguing for trained subjects, who are used to making frequent, analytical, sensory judgments.

4.1. Limitations and future work

A key limitation of this study was the artificial nature of the sample sets: for both the spices and the chocolates, the samples were chosen to span a product category. In a real product-development or other applied situation, it is unlikely that there would be such a structured set of products. Arguably, free linking, which relies on pairwise comparisons, should perform better in these real situations, but this could not be determined from these sample sets. It also remains to be seen whether the lower preference and liking ratings for free linking by subjects will result in lower compliance or lower quality data when the method is used in a non-comparative setting.

The free-linking task also provides some new possibilities for the design of sensory studies. For example, to this point it has not been feasible to conduct free-sorting tests (or indeed projective-mapping tests) in an incomplete-block design, because the sorting space depends simultaneously on all samples. This has restricted the number of samples that can practically be analyzed in a free-sorting study to around 25 actual samples (the number is much higher for visual or text samples). This restriction should not apply to the free-linking task, which is based on a similarity graph of pairwise comparisons, but provides results that are similar or arguably superior to free sorting. Therefore, a logical future study is the investigation by free-linking of similarities in a set of samples large enough to present with an incomplete block design, but small enough to also investigate in full with free sorting in order to determine the comparability of this approach. Incomplete blocks for similarity would be a significant boon to food-sensory researchers in both industry and academia. In addition, given that free sorting appears to become exponentially more fatiguing as the number and sensory complexity of samples increases (see for example Kessinger et al., 2020), it may be hoped that free linking, which requires a larger number of simpler judgments, may perform better with large sample sets, especially when implemented in incomplete blocks as described above.

4.2. Conclusions

In this paper, we present a new, rapid method for assessing similarities among a set of samples: the “free-linking task”. In the free-linking task, subjects are given a set of samples and asked to indicate pairwise similarity according to their own criteria; in effect, as we have demonstrated, subjects are drawing their own individual similarity graph for the samples. The data from free linking can be treated using existing tools for analyzing similarity data, such as DISTATIS, MFA, or even MDS.

The free-linking task explicitly solves two issues with the currently popular free-sorting task: in free sorting, subjects can only indicate one degree of similarity (is/is not similar) and are forced to make fully transitive similarity groups. While previously proposed modifications of sorting like the *hierarchical* and *multiple* free-sorting tasks can solve these respective tasks with replicated or multiple passes of sorting for each sorting, free linking solves both problems at once with only a single task. As we have demonstrated, therefore, the results of free linking provide a more realistic representation of similarity and allow finer and more powerful interpretations than free sorting. However, while the results of free linking are more realistic and robust, the cost is that free linking, because it involves more pairwise comparisons, is also more demanding for the participants. The multidimensionality of free-linking data is also greater, which can be considered either a cost or a benefit, depending on the sensory analyst’s goals. Therefore, we believe that the free-linking

task will be a significant addition to the sensory analyst's arsenal of tools for rapidly assessing similarities, and we expect to see improvements and new uses cases for the tool in the near future.

Acknowledgments

David Orden is partially supported by Project PID2019-104129GB-I00 / AEI / 10.13039/501100011033 of the Spanish Ministry of Science and Innovation. Marino Tejedor-Romero is funded by a predoctoral contract from Universidad de Alcalá. This research did not receive any additional external funding. We would like to thank the volunteers who participated in this study, as well the undergraduate research volunteers at Virginia Tech who helped to coordinate data collection.

References

- Abdi, H., Valentin, D., Chollet, S., & Chrea, C. (2007). Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food Quality and Preference*, 18(4), 627–640. <https://doi.org/10.1016/j.foodqual.2006.09.003>
- Arney, D., & Horton, S. (2014). Network science for graph theorists. In J. Gross, J. Yellen, & P. Zhang (Eds.), *Handbook of graph theory*. CRC Press.
- Beaton, D., Fatt, C. C., & Abdi, H. (2013). DistatisR: DiSTATIS Three Way Metric Multidimensional Scaling (R Package version 1.0) [Computer software]. <http://CRAN.R-project.org/package=DistatisR>.
- Beineke, L. W., Oellermann, O. R., & Pippert, R. E. (2002). The average connectivity of a graph. *Discrete Mathematics*, 252(1–3), 31–45.
- Blanchard, S. J., & Banerji, I. (2016). Evidence-based recommendations for designing free-sorting experiments. *Behavior Research Methods*, 48(4), 1318–1336. <https://doi.org/10.3758/s13428-015-0644-6>
- Blancher, G., Clavier, B., Egoroff, C., Duineveld, K., & Parcon, J. (2012). A method to investigate the stability of a sorting map. *Food Quality and Preference*, 23(1), 36–43. <https://doi.org/10.1016/j.foodqual.2011.06.010>
- Chartrand, G., & Zhang, P. (2014). Distance in Graphs. In J. Gross, J. Yellen, & P. Zhang (Eds.), *Handbook of graph theory*. CRC Press.
- Dehlholm, C. (2015). Free multiple sorting as a sensory profiling technique. In *Rapid Sensory Profiling Techniques* (pp. 187–196). Elsevier. <https://doi.org/10.1533/9781782422587.2.187>.
- Dehlholm, C., Brockhoff, P. B., Meinert, L., Aaslyng, M. D., & Bredie, W. L. P. (2012). Rapid descriptive sensory methods – Comparison of Free Multiple Sorting, Partial Mapping, Napping, Flash Profiling and conventional profiling. *Food Quality and Preference*, 26(2), 267–277. <https://doi.org/10.1016/j.foodqual.2012.02.012>
- Delarue, J. (2015). The use of rapid sensory methods in R&D and research: An introduction. In *Rapid sensory profiling techniques* (pp. 3–25). Elsevier. <https://doi.org/10.1533/9781782422587.1.3>.
- Gross, J., Yellen, J., & Zhang, P. (2014). *Handbook of graph theory*. CRC Press.
- Kessinger, J., Earnhart, G., Hamilton, L., Phetxumphou, K., Neill, C., Stewart, A. C., & Lahne, J. (2020). Exploring perceptions and categorization of Virginia hard ciders through the application of sorting tasks. *Journal of the American Society of Brewing Chemists*, 1–14.
- Koenig, L., Cariou, V., Symoneaux, R., Coulon-Leroy, C., & Vigneau, E. (2021). Additive trees for the categorization of a large number of objects, with bootstrapping strategy for stability assessment. Application to the free sorting of wine odor terms. *Food Quality and Preference*, 89, Article 104137. <https://doi.org/10.1016/j.foodqual.2020.104137>
- Koenig, L., Coulon-Leroy, C., Symoneaux, R., Cariou, V., & Vigneau, E. (2020). Influence of expertise on semantic categorization of wine odors. *Food Quality and Preference*, 83, 103923. <https://doi.org/10.1016/j.foodqual.2020.103923>
- Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical analysis of network data with R*. Springer.
- Lahne, J. (2020). Sorting Backbone Analysis: A network-based method of extracting key actionable information from free-sorting task results. *Food Quality and Preference*, 82, 103870. <https://doi.org/10.1016/j.foodqual.2020.103870>
- Lahne, J., Abdi, H., & Heymann, H. (2018). Rapid sensory profiles with DISTATIS and Barycentric Text Projection: An example with amari, bitter herbal liqueurs. *Food Quality and Preference*, 66, 36–43. <https://doi.org/10.1016/j.foodqual.2018.01.003>
- Orden, D., Fernández-Fernández, E., Rodríguez-Nogales, J. M., & Vila-Crespo, J. (2019). Testing SensoGraph, a geometric approach for fast sensory evaluation. *Food Quality and Preference*, 72, 1–9. <https://doi.org/10.1016/j.foodqual.2018.09.005>
- Orden, D., Fernández-Fernández, E., Tejedor-Romero, M., & Martínez-Moraian, A. (2021). Geometric and statistical techniques for projective mapping of chocolate chip cookies with a large number of consumers. *Food Quality and Preference*, 87, 104068. <https://doi.org/10.1016/j.foodqual.2020.104068>
- Spencer, M., Sage, E., Velez, M., & Guinard, J.-X. (2016). Using single free sorting and multivariate exploratory methods to design a new coffee taster's flavor wheel: design of coffee taster's flavor wheel.... *Journal of Food Science*, 81(12), S2997–S3005. <https://doi.org/10.1111/1750-3841.13555>
- Valentin, D., Chollet, S., Lelièvre, M., & Abdi, H. (2012). Quick and dirty but still pretty good: A review of new descriptive methods in food science. *International Journal of Food Science & Technology*, 47(8), 1563–1578. <https://doi.org/10.1111/j.1365-2621.2012.03022.x>
- Varela, P., & Ares, G. (Eds.). (2014). *Novel techniques in sensory characterization and consumer profiling*. CRC Press. <https://doi.org/10.1201/b16853>.

Distributed Remote Secure E-Voting

” *A bug is never just a mistake. It represents something bigger. An error of thinking. That makes you who you are.*

— **Elliot Anderson**
(Mr. Robot)






The last article presented in this PhD thesis introduces a novel, distributed, remote, verifiable and secure e-voting system, *DiverSEC* [Tej+21]. While there are several proposals already published in the state of the art from the security point of view, the research line remains open, since there are still challenges to address. The objective of this work is to propose a remote e-voting system based on a very cautious threat model, which aims to remove trust requirements from any other actor in the system. At the same time, it aims to offer the security properties which are expected from an e-voting system without risky assumptions and from a ‘security in depth’ perspective, trying to minimize the impact of vulnerabilities.

In order to do so, we distribute different candidates in a network. This creates a conflict of interests which enhances the security of the system. In terms of graphs, this topology defines a directed cyclic graph. We combine this ‘ring’ topology with Shamir’s secret sharing scheme in finite fields and mixnets. As a result, each vote is distributed among the different servers with conflicted interests. Only when they collaborate, they are able to reveal the final vote in a public way, maintaining anonymous the source of the secret shares. The ballot results can be easily calculated and verified by any actor on the system, including voters, which can increase trust in the system.

The following paper offers an in-depth explanation of the e-voting system proposed, and an analysis of its properties, its scalability, the design decisions behind *DiverSEC* and the reasons why these were made.

Article

Distributed Remote E-Voting System Based on Shamir's Secret Sharing Scheme

Marino Tejedor-Romero ^{1,*}, David Orden ^{1,†}, Ivan Marsa-Maestre ^{2,†}, Javier Junquera-Sanchez ^{3,†}
and Jose Manuel Gimenez-Guzman ^{4,†}

¹ Departamento de Física y Matemáticas, Universidad de Alcalá, 28805 Alcalá de Henares, Spain; david.orden@uah.es

² Departamento de Automática, Universidad de Alcalá, 28805 Alcalá de Henares, Spain; ivan.marsa@uah.es

³ Departamento de Ciencias de la Computación, Universidad de Alcalá, 28805 Alcalá de Henares, Spain; javier.junquera@uah.es

⁴ Departamento de Comunicaciones, Universitat Politècnica de València, 46022 Valencia, Spain; jmgimenez@upv.es

* Correspondence: marino.tejedor@uah.es

† These authors contributed equally to this work.

Abstract: A number of e-voting systems have been proposed in the last decades, attracting the interest of the research community. The challenge is far from being fully addressed, especially for remote systems. In this work, we propose DiverSEC, a distributed, remote e-voting system based on Shamir secret sharing, operations in Galois field and *mixnets*, which enables end-to-end vote verification. Parties participate as nodes in the network, protecting their interests and ensuring process integrity due to the conflicting interests. The threat model is very conservative, not letting even the most privileged actors to compromise votes privacy or integrity. Security in depth is implemented, overlapping different mechanisms to offer guarantees even in the most adverse operating conditions. The main contributions of the resulting system are our proposal for secret-sharing among the political parties, which guarantees that no party can compromise the integrity of the ballot without being detected and identified in real time, and the computational and architectural scalability of the proposal, which make it easy to implement.

Keywords: e-voting; remote voting; verifiable voting; secret sharing



Citation: Tejedor-Romero, M.; Orden, D.; Marsa-Maestre, I.; Junquera-Sanchez, J.; Gimenez-Guzman, J.M. Distributed Remote E-Voting System Based on Shamir's Secret Sharing Scheme. *Electronics* **2021**, *10*, 3075. <https://doi.org/10.3390/electronics10243075>

Received: 25 October 2021

Accepted: 7 December 2021

Published: 9 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

During the past decades, the e-voting research field, essential for the development of a bigger concept called e-democracy, has been attracting growing interest. First, because elections are a critical requirement for the proper operation of the current representative-democratic government systems. Second, because we have experienced an explosive digital development that enables alternative voting methods. In this moment, there is powerful and efficient hardware with an acceptable cost. At the same time, there are lots of technology-related and highly-capable professionals. Both circumstances enable combining elections and digital automation.

E-Voting systems present immediate advantages over the paper ballot system. They are more inexpensive and time-efficient, because traditional elections require the orchestration of many people and many resources in order to offer security guarantees. Electronic elections can be not only simpler and cheaper, but also more user-friendly.

This idea is not novel. There are many publications about e-voting, and real-world cases of elections powered using digital resources [1,2]. However, this fact does not imply that e-voting has been properly accepted. These cases have been mostly affected by controversy. In fact, several subsequent analysis over these real-world cases have revealed that the protocols and systems chosen by the officers were not designed from a secure

perspective and had serious vulnerabilities [3,4]. For this reason, we need cryptographic techniques and a secure design. Powerful hardware and a developed IT industry are not sufficient by themselves. A correct design is not only meant to guarantee the integrity and privacy of the vote; it is also important in order to gain the trust of users. Specially nowadays, when the popular opinion of e-voting is not completely favorable.

There are already several e-voting proposals, which use cryptography to solve these requirements, but we can still find weaknesses that must be solved before we take another step forward. This paper contributes to bridge this gap by proposing DiverSEC: a distributed, remote e-voting system (REV) which satisfies desirable properties like ballot secrecy, integrity and verifiability, incorporating elements such as Shamir secret sharing and polynomial rings, and a novel architecture based on a *mixnet*, which distributes responsibility of the process among the different political parties. As a result, none of them can hinder the integrity of the ballot without being detected and identified by the others in real time. The proposal is also flexible and easy to implement, with a computational complexity that is linear to the number of voters, which can be easily divided in electoral districts if needed.

The rest of this paper is organized as follows. Section 2 reviews the desirable properties for an e-voting system, and briefly discusses the most relevant proposals to date. Then Section 3 presents the threat model where our proposal operates, while Section 4 describes the actual proposal: the DiverSEC e-voting system. Section 5 discusses the main properties, strengths and limitations of our approach. Finally, Section 6 summarizes our conclusions and sheds light on some future research directions.

2. State-of-the-Art

2.1. Desirable Properties for E-Voting Systems

There are several properties that are required of e-voting systems, from the basic properties extracted from traditional paper ballot elections (e.g., ballot secrecy or legitimacy), to other requirements, inherent to electronic alternatives. These have been already listed and explained by security researchers [5]. In the following, we briefly discuss the ones more relevant to our work:

- **Ballot secrecy** : Only the voter can know the actual vote. In an e-voting system, this means ensuring vote confidentiality against other voters, against attackers, and even against the very entity performing the election.
- **Legitimacy**: Only previously registered participants can vote. That is, in the final ballot there must not be any vote from an unauthorized participant. In electronic systems, this is guaranteed via authentication.
- **Eligibility**: Just as illegitimate voting is not allowed, legitimate voters can emit only one vote.
- **Accuracy**: The final result accounts for all legitimate votes, without any alteration between emission and tally. This is related to vote integrity.
- **Individual verifiability**: The voter can verify the correctness of the election process she has participated in.
- **Coercion resistance**: A voter can freely emit a vote even under coercion from a third party. The usual way to achieve this property is by means of the impossibility for the voter to prove her own vote. In this way, a third party cannot be sure of the final decision of a voter, which indirectly de-incentivize coercion.
- **Robustness**: The system remains dependable even when some actors actively attempt to corrupt the process. The types of corruption attempts are pre-defined. This is an important requirement for an e-voting system, due to the number of actors involved. The system must be designed taking into account the possibility of corrupted input.

These requirements are not completely independent. Some are connected between them (e.g., accuracy and verifiability). Others have a certain degree of incompatibility, like ballot secrecy against verifiability, and individual end-to-end verifiability against coercion resistance. This is because the ability to track your own vote across the system implies

having exclusive information, and voters could offer their private verification information to an attacker. Therefore, this balance is typically the most difficult to satisfy.

Taking this into account, the goal is to achieve a reasonable compromise, adapted to the current needs of the situation and the voters, and to the advantages that we deem most valuable of the e-voting protocols over traditional ballots.

Besides the basic properties, there are other kinds of concerns about e-voting systems we should solve before deploying them, regarding, for instance, electronic system guarantees. Given any e-voting system, the voters have little knowledge of what's happening under-the-hood, beyond their interaction with the system. This is why an open-source public formalization is deeply encouraged. However, an open-source software implementation of the e-voting system is not enough to solve this distrust, because it does not necessarily imply that everyone in the system is running the open-source implementation [6]. Further on, we will describe the threat and adversarial model we have considered for our proposal but, for the moment, the most cautious choice is to be wary of every actor in the system.

Under this assumption, the only way to guarantee the voter's security and trust, along with the tally's integrity is a secure design. An e-voting system is resistant even to corruption coming from inside the system's most privileged actors only if:

- The voter sends only the minimum information to the system for the entire process to be completed correctly. The most critical point here, is that, with that minimum amount of information, a malicious privileged actor should not be capable of compromising the security of the vote.
- A malicious administration is not able to convincingly fake the tally. A usual way to achieve this are distributed tally systems, in which there are no centralized tally authorities. The degree of distribution may vary, ranging from a few nodes, as is our case, to self-tallying systems where every user may perform the tally independently [7].

These two conditions allow the voter to participate securely in the election without checking the software of the rest of the nodes. [5] regards this property as software independence, although the author also refers to unintentional errors, not only intentional corruption.

2.2. Protocols and E-Voting Systems to Date

We can find a wide variety of e-voting protocols in the literature, for a wide diversity of scenarios, ranging from direct-recording electronic voting machines (DRE) to remote systems, varying the use they make of cryptography, and so forth. The first relevant proposals, driven by renowned cybersecurity researchers are ThreeBallot [8], PunchCard [9], Scantegrity [10]. These three proposals use paper ballots as physical backup and optical sensors to guarantee the integrity and privacy of each user's vote. The proposals are original and creative, and are the context of our state-of-the-art. However, we try to advance through our new proposal, towards a completely virtual, remote and verifiable approach.

Regarding the use of cryptography, we can find several categories of protocols, depending on the mechanisms used [5]:

- **Mixnet** [11]: They use a layered architecture where a chain of proxy servers is created. These servers receive an input with data from different sources, and they pass the data to the next proxy after a random permutation. This makes impossible to trace the source of an input to the system (unless all nodes are corrupt). This is the underlying principle of networks like Tor [12]. Some protocols that use *mixnets* are [13,14].
- **Blind signature** [15]: With a blind signature a principal obtains a cryptographic validation for a message from another principal, without disclosing the message to the latter. Typically, a voter gets the blind signature for her vote from the administration, without the administration knowing the content of her vote. Some protocols that use blind signatures are [16,17].
- **Homomorphic encryption** [18]: They use cryptographic techniques enabling to process the information without decrypting it, therefore handling votes in an aggregate manner without revealing individual votes [7,19,20].

- **Blockchain [21]:** Blockchain is an emerging technology that jumped to the popular culture during the last years, mainly because it is the base of cryptocurrencies. The principle of blockchain is a data structure that organizes items of information in blocks, which are connected through hashes. The blockchain data is distributed, stored and validated by the nodes. This is just a very small summary of this technique, since it is the most complex point in this list, with lots of variations, different consensus algorithms, and further analysis.

We have seen an extraordinary number of e-voting proposals using Blockchain [22]. However, we consider that this technology is not suitable for a really secure e-voting system under our current threat model (at least in its current state). The main drawbacks of the blockchain technology are its poor scalability, and the needless computational and energetic waste [23]. Recently, the alternative ‘Proof of Stake’ consensus algorithm has been proposed to replace ‘Proof of Work’, as a means to solve the computational problem, however, it still presents serious security drawbacks [24]. Although we have evaluated that, in our threat model, its advantages do not overweight its problems, there are some key design decisions we would like to extract and apply to our proposal: a distributed network, and global verifiability of the messages sent through hashes.

- **Shamir’s secret sharing [25]:** It is a well-known cryptographic technique, although it is used significantly less often in e-voting [26]. It has been recently used in very diverse application areas, such as randomness [27] and authentication [28]. The naive version of Shamir’s secret sharing presents some known flaws, but, at the same time, there are known solutions to these problems. For example:
 - If this scheme is deployed using standard integer arithmetic, it does not fulfill all the basic properties of the secret sharing concept. This is because shares leak data about the secret. This is solved easily in our proposal using a finite field.
 - This scheme is not verifiable. While there are schemes that aim to solve this issue, like the ones based on PVSS [27], in our proposal we solve it using our own integrated mechanism, as we will see below.
 - If the shares are revealed sequentially, the last shareholder can alter the last share in order to manipulate the output of the interpolation process. The easiest way to overcome this issue is to force the shareholders to publish a hash of their shares first, so they cannot change their share. This is typically called ‘commitment’.

These groups are not exclusionary. In some cases, multiple techniques are overlapping in the same e-voting system, and approaches are classified according to the main principle of the schema.

In this moment, the most accepted and relevant remote and verifiable proposals are Helios and Civitas/JCJ:

- **Helios:** Helios [6] is an e-voting system designed to be used through a web browser. It takes Benaloh’s *Simple Verifiable Voting* [29] as base. It is a centralized system that attempts to provide the two most basic and fundamental properties, integrity and privacy. The author prioritizes absolute integrity over the secret ballot in the worst case. Helios deliberately ignores the problem of coercion, arguing that it is an inherent problem of REV proposals. The verification processes are interactive, that is, the voter’s guarantees do not depend on the amount of information sent, or the information publicly available. Instead, the verification is based on the outcome of the voter’s interaction with the central Helios server and/or with the different participants of the *mixnet*.
- **Civitas/JCJ:** Civitas/JCJ is another approach to remote voting, which employs anti-coercion measures, through duplicate identities [30,31]. During the registration phase, a valid ID is obtained, in addition to other fake credentials. These fake credentials, which must be removed from the final count, are able to confuse a potential attacker as they are indistinguishable from the real one. In order to satisfy this complicated requirement, this system relies on loose assumptions. This is, voters have to place

more blind trust because its threat model is more permissive. Under a more critical analysis, these more relaxed assumptions are directly against the property discussed previously: strong and blind voter guarantees. The initialization phase is distributed, while vote counting and publication is done from the central server. Like Helios, vote privacy is based on a *mixnet* that obfuscates the origin of each virtual ballot, encrypted by the user.

Additionally, the most recent research has tried to refine both proposals. For example, we can mention:

- **Ordinos** [32]: Ordinos is an improvement to the Helios voting system. Its most prominent feature is called 'tally-hiding'. Through the use of a multi-party computation network, the system exposes a result calculated from the tally, instead of the tally itself. Like Helios, it is verifiable end-to-end, accountable, and it trusts the central Helios server.
- **Incoercible fully-remote electronic voting protocol** [33]: This system is not exactly based on JCJ/Civitas. It takes it as an inspiration, while the authors address its recognized problems and propose a new alternative that overcome said vulnerabilities. However, we can still perceive relaxed assumptions.

3. Adversarial and Threat Model

In any system, from a security point of view, it is crucial to specify the participants, their motivations, and what malicious behaviour is expected from them. Once we have established the desirable properties from our system and the threat model, we can analyze the proposed system, its strengths and vulnerabilities.

In our e-voting proposal, there are three different types of participants.

- **Administration:** It is the organization that manages the entire voting process. It would correspond, for example, to the government in the case of a national or regional election. We can expect the administration to be dependent on one or more political parties in power at the time of the election. Their main interest is that the voting process develops correctly and that the results are valid, since all users can verify their vote. Moreover, considering that this organization may not be independent, we have to consider a malicious partisan interest. It could try to manipulate the result in its favor with its special privileges.
- **Parties:** Each political party participates in the electoral process with a server that collects part of the fragments of information needed to later reconstruct the result. It is to be expected that the parties will have conflicting interests, for selfish reasons in a competitive environment, and for differences in their doctrines. Perhaps, in some cases, there may be complicity between two parties close in the ideological spectrum, although experience reveals that competition for votes is the priority during elections, unless these two formations present a joint candidacy (in which case, for technical purposes in this model, they are counted as a single party). Joint or independent corruption of the total number of parties is not considered viable. Their main motivation is to win as many votes as possible. In the balance between the struggle of interests from all parties, the security of the vote is ensured. Their possible attacks are based on discovering the vote of some users, buying their votes, or modifying the votes during the voting as much as possible so that the result is favorable to them. The rest of the parties will not accept attacks, since their interest is also to protect their votes.
- **Voters:** They are the users of the system, registered in a census, who have the right to vote (only once) in the voting system. They are the least privileged participants within the system and they do not hold responsibilities. This is the largest group, and at the same time the most varied and unpredictable. The margin of action they may have is reduced to a minimum.

Their primary legitimate interest is to vote, and then verify their participation in the final tally. However, malicious voters may try to boycott the election or vote more than once to gain an illegal advantage for their political choice.

It is important to emphasize the fact that political parties and the administration are public entities, and are subject to public scrutiny. This means that there are two levels of security measures against attacks coming from the administration or from political parties (which are the privileged agents of the system, since voters barely have attack margin). At the same time, we can actively try to block attacks and try to make them visible. The moment one of these entities attempts an attack to manipulate the vote, it will be detected by the rest of the actors. The parties and the administration must maintain a positive image in the eyes of the citizens. The empirical and demonstrated exposure of an attack coming from a public entity would terribly undermine their credibility.

4. The DiverSEC Proposal

Here we describe our remote e-voting proposal in detail, with a section for each of its stages: initialization, vote generation, authentication and casting, *mixnet*, recomposition and publication.

4.1. Initialization

In the first phase of the protocol, the administration establishes all the parameters and publishes all the information to the rest of the participants. We will list all the actors involved, the parameters that are published and their notation.

- **Parties** $\{p_1, p_2, p_3, \dots\}$: P parties participate in the voting process, noted as $p_i, 1 \leq i \leq P$. Their network addresses IP_{p_i} and all their public keys (that have been generated for this system) K_i are published. Every party has been assigned a public coordinate in the X axis. This parameter is important for the purposes of Shamir's secret sharing scheme. From now on, and without loss of generality, we will assume that for every $p_i, x = i$.
- **Options** $\{0, 1, 2, 3, \dots\}$: There are O options to vote on. This number of options is necessarily a prime number. If the number of options required by the situation is not a prime number, then the next higher prime is selected by adding blank padding votes. The options include the blank vote and all parties as a minimum. Additionally, it is possible to add other alternative options. The list, known at initialization time, includes the numbers in \mathbb{F}_O along with the option each one of them represents. For example, this is a possible list for $O = 7$:
 - 0: Blank vote
 - 1: Party A
 - 2: Party B
 - 3: Party C
 - 4: Party D
 - 5: Alternative E
 - 6: Blank vote
- **Voters** $\{v_1, v_2, v_3, \dots\}$: Citizens who have the right to cast a single vote and to verify the result later. They are identified by their public key, within the current PKI. The census is public, as it happens in a traditional election, linking voter and public key. V is the number of registered voters, and $v_j, 1 \leq j \leq V$ is the denomination of each voter. Each voter's public key is $K_{pub,j}$.
- **Mixnet sequence (M)**: The administration establishes an order for the parties, or in other words, a distribution of the set $\{1, 2, \dots, P\}$. Any permutation is equally valid, but we assume from now on that an ascending numerical order is chosen for its simplicity. M is, in fact, a cycle, since the last party of the sequence will precede the first. We denote by M_i the sequence M shifted so that the last element is i .

- **Authentication challenge (R_{au}):** Random prefix that will serve as a challenge in the authentication phase, to avoid replay attacks.
- **Recomposition challenge (R_{rc}):** Random prefix that will serve as part of the random number that will later be used to recompose the fragments of each vote.

All these parameters must be communicated to each voter, personally or through a public ‘board’. The method used is irrelevant. From now on these elements are assumed to be common knowledge. Figures and examples will be used to illustrate the procedure graphically, with four parties and five options, hence $P = 4, O = 5, M = \{1, 2, 3, 4\}$.

4.2. Vote Generation

In order to cast a vote, each voter must generate a polynomial that reflects her or her voting intention. This phase is divided in two parts, the generation of the polynomial and the packaging.

We work on the finite field defined by the elements of $\mathbb{F}_O, \text{ mod } O$. The voter generates a random polynomial of degree $P - 1$, $f(x) = a_0 + a_1x + a_2x^2 + \dots + a_{(P-1)}x^{(P-1)}$, whose threshold is $k = P$. The free term a_0 represents the option chosen by the voter, according to the published list of options. The rest of the terms $\{a_1, a_2, \dots, a_{(P-1)}\}$ are random integers in the finite field \mathbb{F}_O . Fixed the polynomial $f(x)$, all the shares $f(i)$ are calculated, one for each party, $1 \leq i \leq P$. An example for the polynomial $f(x) = 2 + 3x + x^2 + 3x^3$ is illustrated in Figure 1.

$$f(0) = 2; f(1) = 2 + 3 + 1 + 3 = 4; f(2) = 2 + 1 + 4 + 4 = 1; \dots$$

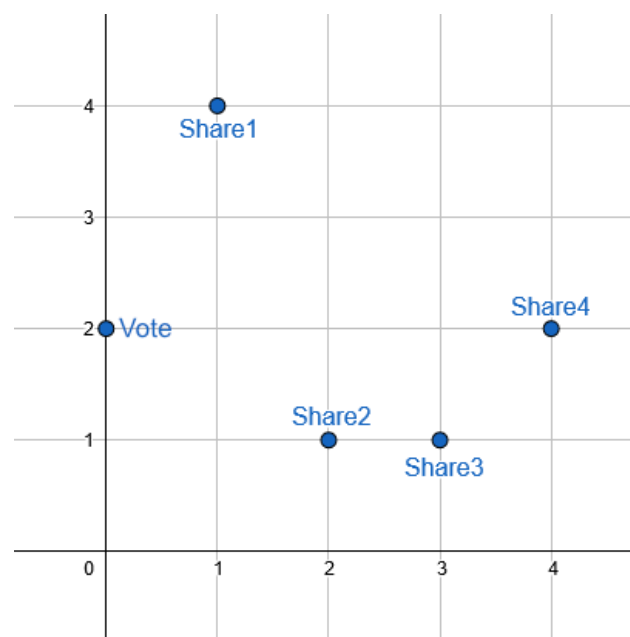


Figure 1. Obtaining shares from the generated polynomial.

With the computed shares, the voter generates a recomposition nonce using the prefix and a random private part generated by each voter j : $N_{rc,j} = R_{rc} \parallel Rand_j$. The voter must remember her $N_{rc,j}$ in order to verify her vote later. Then, the shares are encoded and encrypted as follows.

1. The share and the nonce are concatenated: “ $f(i) = \dots, N_{rc} = \dots$ ”
2. The shifted *mixnet* sequence M_i is obtained, so we can iterate from the last element i to the first.
3. For each index m , the voter takes the current state of the share and the nonce and encrypts it for the party p_m with its public key.

4. During each iteration, the voter saves a hash of the current state of the packet of share and nonce, from plaintext until it has been encrypted P times.

After this process, each point $f(i)$ of the polynomial attached to N_{rc} has been encrypted using the public keys of all parties. The encryption order is determined by each shifted sequence M_i . The result is called SC_i . In our example scenario with four parties, the sequence M_2 used to encrypt the second share, would be:

$$M = \{1, 2, 3, 4\}, M_2 = \{3, 4, 1, 2\}$$

$$SC_2 = \left\{ \left\{ \left\{ \left\{ f(2) = 1, N_{rc,j} = d5a3bb \dots \right\}_{K_2} \right\}_{K_1} \right\}_{K_4} \right\}_{K_3},$$

where $N_{rc} = d5a3bb \dots$ represents the random nonce.

The voter must store an array of size $P(P + 1)$. This data will be used to verify the correct transmission and decryption of each share through the network. Figure 2 shows an example for a voter Alice.

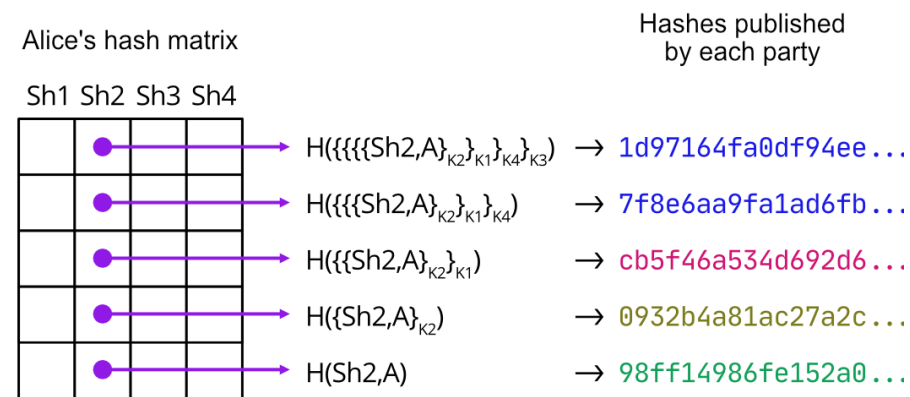


Figure 2. Hashes corresponding to voter Alice, for *mixnet* stage verification.

4.3. Authentication and Casting

Authentication is performed via public key, since the voter census includes their public keys and has been disclosed in the first phase. A process analogous to the public-key method of the SSHv2 protocol, documented in RFC4252 [34], is used. In this system, the voter digitally signs her intention to vote, along with her name, her public key, and the challenge R_{au} . This signature is sent to each of the parties and the administration.

Once the voter is authenticated, parties then accept only one encrypted share from each voter. The voter distributes her information as follows. Each point $f(i)$ must reach the party p_i , which is in charge of the position $x = i$. For this reason, the voter takes each encrypted share SC_i and observes its order M_i , which ends at the party p_i . The first element of M_i is the start of the chain, so the voter must send each SC_i to the first index of that list (Figure 3).

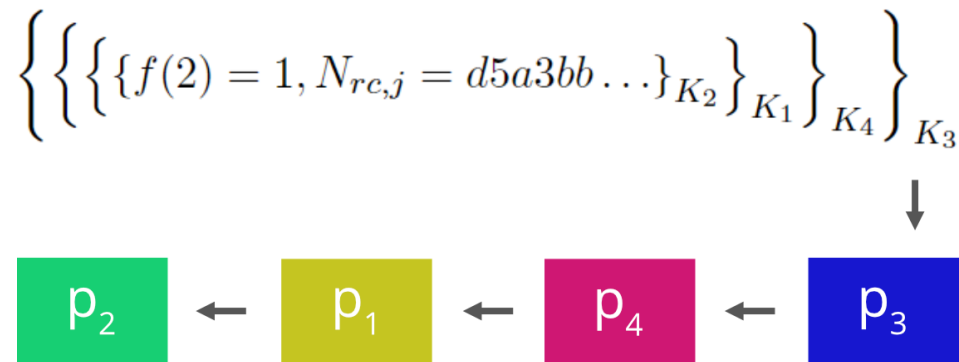


Figure 3. Each SC_i traveling through the node chain.

4.4. Mixnet

The *mixnet* phase begins when all voters have cast their shares. Of the total census, of size V , only a subset consisting of V' voters, where $V' \leq V$, participates. It is to be expected that not all voters use their right to vote. This subset is known, moreover, given the authentication signatures, which are shared by all parties and the administration.

Before starting, all parties confirm the reception of V' encrypted shares and different voter signatures. They publish a list of hash of all the SC_i they own, for voter verification. From now on we note as a hash-list the result of calculating the hash of each of the encrypted shares held by a party at a given time, in ascending order with respect to the hash, instead of the origin. This detail is critical; the order of the hash-list does not provide any information about the order of the preimages. Then, the rounds of the *mixnet* begin:

- Each party calculates the hash-list of all the encrypted shares received, and compares it with the hash-list published by the previous node, to make sure that there has been no manipulation. The first round does not rely on this check, since the previous node is not a party but the voters, and they have already checked the hash-list at the beginning.
- Each party decrypts all shares received with its private key, removing one of the encryption layers.
- Each party calculates the hash-list after decryption, and publishes it for verification by the voter the following party.
- Unless it is the last round (the last encryption layer has been removed), each party randomizes the order of its encrypted shares, and sends them to the next node, according to the cyclic order defined by M .

After P rounds, every encryption layer is removed, and each party p_i has V' shares in plaintext. Figure 4 shows this process by isolating the shares of party p_2 . Five voters cast their SC_2 to p_3 , and reach their destination after passing through all parties. Each one decrypts the layer it can, publishes the hash-list and transmits their data to the next party, randomizing its order. In each published hash-list, an entry has been highlighted. Figure 2 shows a column of the matrix that Alice has saved in a previous phase. Alice can verify that, in each of the published lists, her share is in the state it should be. As a result, her share has not been tampered with or deleted.

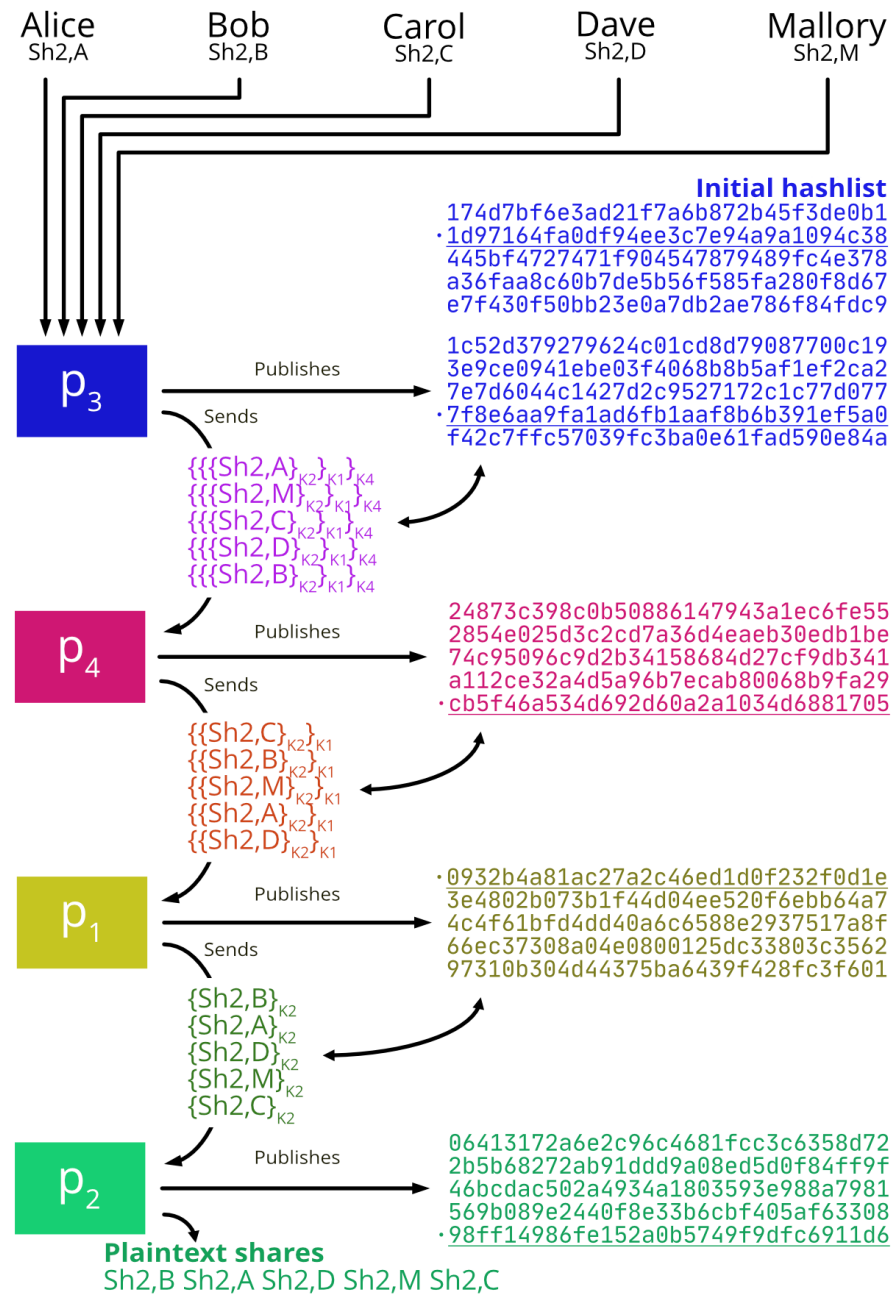


Figure 4. Mixnet example isolating shares from p_2 .

4.5. Recomposition

Each party has, at this point, a list of V' nonces $N_{rc,j}$. Voters themselves may attempt to attack the integrity of the election. Therefore, the list of nonces for each party is published. Those entries that do not include the required prefix R_{rc} are removed, and those that do not appear on all lists are removed. The rest, which are valid, are sorted according to a common rule. For simplicity, we have assumed an ascending order.

In Figure 5, we show how the nonces for our example have been ordered, and one inconsistent nonce, sent by Mallory, has been removed. Voters know their nonce, but no one else can link nonce to voter.

	p ₁	p ₂	p ₃	p ₄
0: Bob	00f6bc	00f6bc	00f6bc	00f6bc
1: Dave	8b7785	8b7785	8b7785	8b7785
2: Alice	92d58d	92d58d	92d58d	92d58d
3: Carol	cbe568	cbe568	cbe568	cbe568
Mallory	e556e7			
		fe1d89		
			dd3451	
				da4a56

Figure 5. Ordering of the nonces.

From now on, we work on the finite field $\mathbb{F}_{O^{V'}}$. Being a finite field which cardinality is the power of a prime number, the elements of the set are not numbers anymore, they are polynomials, according to the theory of finite fields [35]. In this case, the polynomials that form this field are of degree $V' - 1$ and the coefficients are integers mod O . We will call them sum polynomials $s(\alpha) = t_0 + t_1\alpha + t_2\alpha^2 + \dots + t_{(V'-1)}\alpha^{(V'-1)}$. We will denote as $s_i(\alpha)$ the sum polynomial of each party p_i . Each of them composes its sum polynomial as follows: Taking the ordered list, each party $p(i)$ determines all coefficients $(t_0, t_1, t_2, \dots, t_{(V'-1)})$ with all points $f(i)$ received, ordered according to their nonce. In other words, t_0 will correspond to the $f(i)$ whose nonce is first in the list, t_1 will correspond to the $f(i)$ whose nonce is second in the list, and so on.

4.6. Publication

After computing the polynomial sum $s_i(\alpha)$, each party publishes the hash digest of that polynomial, using random padding to avoid precalculation attacks. The space of possible polynomials is $O^{V'}$. Therefore, given a large number of choices and voters who have participated, an attack based on the precalculation of all $H(s(\alpha))$ may be unapproachable. However, adding random padding provides extra security without drawbacks.

Once every polynomial hash $H(s_i(\alpha))$ has been published, the parties publish the sum polynomial in plaintext. It is possible to check that the hash and the polynomial match, and in this way we avoid asynchronous publication attacks, this is, that the last of the parties can observe the information of the rest to elaborate an illegal $s_i(\alpha)$, which does not correspond to reality, in order to tamper the results. In Figure 6 we see an example of sum polynomials.

The voting result can be checked by all actors in the system at this point. We denote as $u(i)$ the polynomial of degree 0 whose free term is i . By interpolation, we must compute the polynomial $R(x)$, given the points $\{R(u(i)) = s_i(\alpha), 1 \leq i \leq P\}$. The result of the voting process is found in $R(0)$, which is a polynomial of degree $V' - 1$, with $V' - 1$ coefficients. Each of the coefficients, integers of the field \mathbb{F}_O , is a vote, according to the table of options defined from the initialization. According to the example in the figures, we look for a $R(x)$ such that:

$$R(1 + 0\alpha + 0\alpha^2 + 0\alpha^3) = 3 + 1\alpha + 4\alpha^2 + 2\alpha^3$$

$$R(2 + 0\alpha + 0\alpha^2 + 0\alpha^3) = 1 + 0\alpha + 1\alpha^2 + 2\alpha^3$$

$$R(3 + 0\alpha + 0\alpha^2 + 0\alpha^3) = 4 + 1\alpha + 1\alpha^2 + 3\alpha^3$$

$$R(4 + 0\alpha + 0\alpha^2 + 0\alpha^3) = 3 + 2\alpha + 2\alpha^2 + 2\alpha^3$$

By interpolation, the polynomial of degree 3 we are looking for is:

$$\begin{aligned} &(4 + 1\alpha + 2\alpha^2 + 1\alpha^3) \\ &+ (4 + 4\alpha + 3\alpha^2 + 3\alpha^3)x \\ &+ (4 + 3\alpha + 1\alpha^2 + 1\alpha^3)x^2 \\ &+ (1 + 3\alpha + 3\alpha^2 + 2\alpha^3)x^3 \end{aligned}$$

The result $R(0) = (4 + 1\alpha + 2\alpha^2 + 1\alpha^3)$ reveals that option 1 has obtained two votes, option 2 has obtained one vote, and option 4 has obtained one vote.

The voter can now verify her vote in $R(0)$. First, the voter locates the nonce within the ordered list. If this position is c , the coefficient of the term α^c of the polynomial $R(0)$ must match the voter's option, that is, the a_0 of the polynomial it generated at the time. In addition, the coefficients of the term α^c of the sum polynomials $s_i(\alpha)$ must match each share sent to each party, $f(i)$. In Alice's case, she must find the coefficients of α^2 in the sum polynomials $s_i(\alpha)$ and in the result $R(0)$.

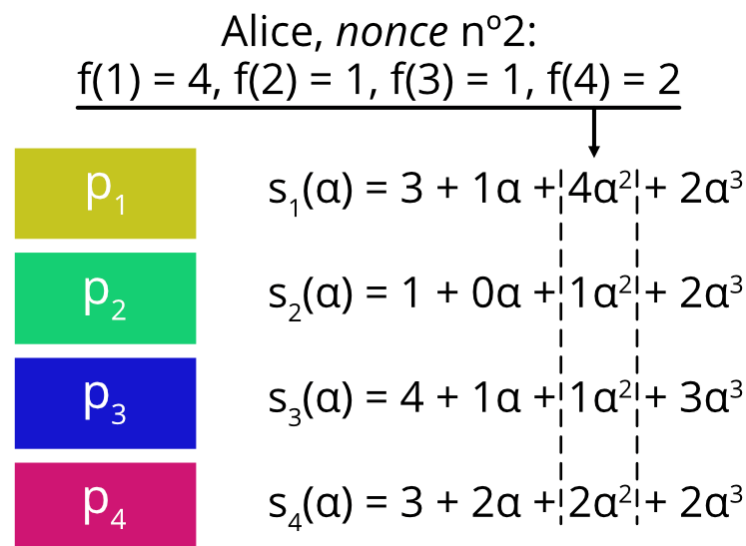


Figure 6. Sum polynomials according to nonces ordering.

5. Results and Discussion of the Contributions

In this section we describe the main results of this work and the ways in which it contributes to the existing state of the art. First, we review the properties achieved by the proposed schema. Then we outline its main strength, and also its current limitations. Finally, we perform a computational analysis of the system.

5.1. Properties Achieved

In Section 2.1, we enumerated the list of properties expected from an e-voting system according to the literature. It is appropriate to review them and check which ones this proposal satisfies and under which assumptions.

- **Ballot secrecy:** In a REV system, this property takes on a double dimension. No participant in the network, which is shared by voters, administration and parties, should know the content of a voter's ballot. We must decouple the vote from the voter's identity. This includes the data used to authenticate the voter, but also from any other data which could be traced to a specific voter, such as its IP address. The system must provide a solution to this problem without telling the voter to use a proxy or hidden network service.

The parties are directly involved in the process, collecting the shares, decrypting them layer by layer and sending them to the next node, and then publishing the sum. For this reason, the parties could violate the ballot secrecy, in two ways:

1. Decrypting each voter's vote as it is cast. It seems impossible, if we consider the cipher used to be secure. A single party must know or break the keys of all the other parties and collect every share. Obviously, if the encryption is broken or the keys are violated, there is no privacy.
2. Linking a clear vote to its origin. At the end of the *mixnet* phase, each party receives a point of the polynomial function along with a nonce. If a party is able to identify its origin, either by its IP address or by association with the digital signature that authenticates it, then it would violate the secret ballot. However, to identify the origin of a nonce from the beginning to the end of the *mixnet*, it is necessary that all the participating parties, without exception, are corrupt and conniving with each other. If only one of the parties randomizes the shares without revealing the order, it is impossible to trace any share to its source.

Taking this into account, we claim the system satisfies ballot secrecy under our threat model.

- **Integrity:** The integrity of each vote depends on the *mixnet* phase. Once the nonces and sum polynomials are published, the outcome of the election is determined. However, up to that point, the shares of each user are left vulnerable during the *mixnet* phase. During one of the iterations, a party could alter the shares it owns. The loss of the integrity of one or more shares of each voter results in the elimination of that vote (since the nonces will not match). The corrupt party in question does not know which vote it is manipulating, because of the properties of Shamir's secret sharing. For this reason, it would not be able to predict the outcome either. Assuming he could change one vote for another, the result would be random and there would be no incentive to perform such an attack.

In addition, there are two mechanisms to detect this phenomenon. On the one hand, users can check the hash-list that is published per party and per round during the *mixnet* phase. In this way, the voter is able to identify where an attack has occurred, in which round and which party was the culprit. On the other hand, the parties are forced to send consistent data. This is because, by randomizing and forwarding the shares from party A to party B, party B can check, before decrypting, that the hashes of what it receives correspond to those published by party A. With these two mechanisms, corruption is detected both in the transmission between parties and in the decryption, so there is no way to violate the integrity of the vote without it being detected by the voter.

- **Legitimacy and eligibility:** The total census of voters, with their identity and public key, is known from the beginning of the election process. This detail, coupled with the fact that the proposed system is distributed, prevents a successful attack in which an illegitimate voter participates, or a legitimate voter casts two votes. The length of the signature list and the hash-lists must match for all parties, and must be constant throughout the iterations.
- **Verifiability:** Voters can verify their votes end-to-end against the final tally thanks to the reconstruction challenge. The shares published must be the ones cast, and the vote must match the one cast. In addition, each voter can also verify their hashes during the *mixnet*. This verification reveals which party is the culprit of the attack.
- **Coercion resistance:** This proposal does not implement receipts issued by the administration or by any other means to verify the vote. Instead, it is the voter who generates a random nonce that identifies all the shares. Once these nonces are published and sorted, then there is no possibility of a coercion attack. Prior to this point, however, there is a vulnerability by which a voter could demonstrate her or her vote to a third party. If this voter communicates the nonce used to a third entity, this entity can check whether this nonce is in the ordered list of nonces, and the associated vote.

- **Error tolerance:** There are several expected error cases in this proposed voting system. Although not all of them have an associated troubleshooting protocol, they are detectable in real time. The following cases are considered, ordered according to the phase in which they might occur:
 - Voters using integers outside the associated finite field.
 - Voters sending authentication signatures and/or signatures only to part of the parties.
 - Parties refusing to accept authentication signatures and/or shares.
 - Parties that tamper with shares during the *mixnet* phase.
 - Parties that publish points and/or nonces that do not correspond to what they receive.

5.2. Computational Analysis

The proposed voting system has three main parameters on which the computational complexity depends: the number of parties P , the number of choices O and the number of voters V .

- P : The number of parties determines the number of iterations of the *mixnet* phase. In addition, the voters have to encrypt their P shares, P times each share. The maximum complexity observed with respect to P is quadratic, $O(P^2)$, during the second phase (vote generation), as we will observe in the following analysis.
- O : The number of options determines the finite field in which the users will operate in order to generate their vote and the finite field in which the voting result will be interpolated, \mathbb{F}_O and \mathbb{F}_{OV} , respectively. It does not affect the number of operations done by the actors in our system.
- V : Assuming that all voters participate (the most conservative case), the number of voters affects the finite field \mathbb{F}_{OV} at the time of polynomial interpolation. The voting result is a polynomial with as many terms as voters, as are all polynomials published by the parties. Thanks to the optimization presented in the following analysis, the maximum complexity for this dimension is linear, $O(V)$, found in the last three stages of the protocol.

In order to analyze the feasibility of the system, it is necessary to delimit the expected range of values for each of these parameters. The reality is that in each country, the configuration P , O , and V may change.

- P : The major parties tend to vary within the interval $[2, 10]$, but the total set of parties can be expected to be significantly larger. In any case, it is estimated that it will never exceed 50 parties, and that this is a stable and controlled number.
- O : For simplicity we can assume that the value of O equals $P + 1$. In the most conservative case, all political parties will actively participate.
- V : This is the most critical value for the computation analysis. In the worst case, we can say that there are as many voters as the population of legal age with the corresponding nationality. However, this assumption is rather extreme, since the most typical case will be to fragment the elections in different polling stations, by necessity of the corresponding electoral law or for technical convenience. In any case, it is undoubtedly the largest parameter, and can be expected to be at least 1000 individuals.

Finally, we can list the phases of the system with their computational complexity. For better scalability, we look for a low complexity class, so we can increase the number of voters without making the execution of the process unfeasible. Typically, linear complexity is usually desired.

1. **Initialization:** During the initial phase, the configuration is established along with all the necessary elements, including addresses and public keys, are made public. This phase does not require computational power.
2. **Vote generation:** Each voter generates a polynomial, computes the different shares and packs them for distribution. The generation of a polynomial involves generating

$P - 1$ random coefficients, and computing P points of the corresponding polynomial function. These two operations are independent, so a linear complexity is observed $O(P)$. The second part involves encrypting each of these points for all parties. That is, each of these P points is encrypted P times. The complexity in this case is quadratic, $O(P^2)$.

3. **Authentication and distribution:** Each voter distributes authentication signatures and her or her shares. The complexity is linear $O(P)$.
4. ***mixnet*:** The mixing network involves as many rounds as there are layers of encryption applied to the shares, therefore, the number of rounds equals P . Each round is independent. The operations that occur in each round are decryption, randomization, and computation of the hash of all the shares they manage. The complexity is linear in these two dimensions, $O(P)$ and $O(V)$.
5. **Reconstruction:** The reconstruction of the polynomial sum of each party is, in short, to order some coefficients according to the consensus list. The complexity is linear with respect to the number of voters $O(V)$.
6. **Disclosure:** This is the most computationally complex phase. Anyone can interpolate the polynomial of degree $P - 1$ into the finite field \mathbb{F}_{O^V} that fits the values provided by the parties. The complexity of the polynomial interpolation itself depends on the algorithm used, especially if we take into account that we do not need the whole polynomial, but the independent term. Shamir himself states in her proposed scheme for shared secret [25] that there are algorithms of complexity $O(P \log^2(P))$ for polynomial interpolation. On the other hand the concern about operations on the finite field \mathbb{F}_{O^V} can also be solved, by translating each operation with polynomials of degree $V - 1$ to V independent operations on the field \mathbb{F}_O , whose elements are integers mod O . In this sense, the complexity is linear $O(V)$.

The computational analysis reveals that the highest complexity observed is quadratic, regarding the number of parties in the vote generation phase, when encrypting the points obtained. This was to be expected, since the voter must obtain a matrix of hashes for subsequent verification. However, the number of parties is going to remain small, and this is not expected to be any problem for any modern CPU.

The other problematic aspect is the use of operations with very high degree polynomials, $V - 1$. The parameter that can grow the most is that of voters. In fact, a high number of these are expected. However, the complexity is linear with respect to this parameter thanks to the ability to independently interpolate each of the coefficients of the voting result.

5.3. Strengths and Limitations of the Proposal

The main strengths of the DiverSEC system can be summarized as follows:

1. The addition of secret sharing implies that the system is distributed. Distributed systems have added complexity, but that complexity creates, in return, a security tension between nodes, where parties cannot perform attacks without being detected by each other. In a centralized model, the different voting authorities could perform attacks due to their privileges, and there are only two options against this approach: prevent the attacks or trust the authorities. After distributing those privileges among independent actors, these attacks are no longer possible. In our case, each *mixnet* node knows just one part of the information. For this reason, their attack surface is greatly reduced.
2. The privileges of this system are not only distributed, they are distributed among entities with clearly competing interests. For this reason, they will make sure to report attacks from the rest of the nodes. On top of that, political parties are public entities, subject to scrutiny and judgment by the voters.
3. Attacks are prevented twofold. First, attacks are detected in real time, thanks to the verification processes. Both political parties and voters are checking the hashes in every step of the *mixnet*. It is impossible to tamper a vote without being detected as long as the chosen hash function remains secure against collisions. Second, it

is impossible to perform targeted attacks. During the *mixnet* phase, the shares are encrypted, so the nodes do not know the content nor the voter of any share. Replacing the content with a new generated share is not possible, since the attacker ignores the recomposition nonce. Even if we ignore these facts, in the Shamir's Secret Sharing Scheme, it is not possible to craft a malicious share that gives a desired arbitrary output (as long as we use a commitment scheme to publish the shares).

4. In this case, verification is based on a very simple process, based on public information. Each voter can verify the processing of her shares by comparing her stored hashes against the list published by each node during the *mixnet*, and can verify her vote end-to-end by identifying her private nonce and the matching plaintext shares and secret. This verification is not interactive and is persistent over time.

It is also important to mention that the proposed scheme is easy to implement. With the computational complexity analysis described above, we can see that a current standard server machine is able to process all the information, given the operations that we have listed until now. Any current end-user device is enough for the voter tasks, such as generation, distribution, and verification. The computational complexities are linear for the number of voters (which is the clearly the higher dimension of a voting system) and quadratic for the number of parties. However, if the amount of voters grew to be a problem, this scheme permits easily distributing the election. This approach is similar to the current election system, which distributes the complete census in different electoral districts.

One of the current concerns in designing crypto-based systems is the impact that the advance of quantum computing may have in them, so this is an issue that must be taken in consideration when proposing a new system [36]. As far as our e-voting system is concerned, the Shamir secret sharing scheme should not be impacted by the advances on quantum computing. Traditional asymmetric cryptography primitives such as RSA, however, are expected to be threatened by these advances. Fortunately, our system is flexible in the use of cryptographic primitives, so it does not need any specific primitive for encryption or signing, and thus could be easily adapted to any quantum-safe standard used at the time of implementation. Likewise, the public key management mechanisms or bulletin board implementations (i.e., [20] or [37]) are not specified either.

Regarding limitations, the main problem with this proposal is the ability of the voters to prove their vote accurately to a third party, in the lapse of time between the generation of the vote and its publication of the tally. This enables voluntary vote buying, or forced coercion.

Secondly, the proposal has no real-time troubleshooting protocols to correct integrity attacks. The current proposal can only detect them, and expose the attacker.

Finally, there are some aspects of an actual e-voting system which are out of the scope of our proposal. For asymmetric encryption to be usable in practice, a PKI infrastructure must be in place for citizens and political parties. In a similar way, actual device protection is out of the scope of our work, but there are plenty of security mechanisms that could be used to protect the computation if needed, such as security enclaves like the ones used in [36]).

6. Conclusions and Future Work

In this work, we propose a novel distributed, remote e-voting system (REV) for elections. Our proposal involves three novel elements in addition to the techniques typically used in the state of the art:

1. Shamir's secret sharing scheme. Its use has generally been limited because of its integrity vulnerabilities. Only a very recent proposal [26] uses it, but it solves this problem in a completely different way: it employs DRE machines under the assumption that those machines are secure.
2. Galois' field. Shamir's secret sharing needs finite field operations in order to eliminate one of its vulnerabilities, but it has not previously been exploited in order to limit the possible values to vote on. The size of the field was traditionally irrelevant, and a

sufficiently large prime number was chosen. In this case, in our field $\mathbb{F}_{QV'}$ neither the base nor the exponent are arbitrary, but correspond to voting parameters.

3. The proposal of a distributed system in which the parties themselves form the voting network. Since we are confronting the selfish interests of each political party against the rest, this creates a positive balance that empowers our security model. It is the electronic equivalent of having supervisors from every party inspecting the ballot boxes in person.

The designed system satisfies the main concerns and needs of REV systems. In an exhaustive review of the list of desirable properties, this system fulfills all but coercion resistance, which is one of the biggest problem in the current state-of-the-art.

The two most critical dimensions, ballot privacy and integrity are covered in depth. It is to be expected that vulnerabilities will arise that undermine the security offered in any protocol, system, or cryptographic primitive. For this reason, multi-layer security provides a shield against these undiscovered vulnerabilities. Encryption protects confidentiality, but at the same time, each vote is divided into different shares, which do not reveal information about the secret unless a single entity receives every single share in plaintext. Publication of the verification hashes in turn protects integrity and, at the same time, removes the incentive for parties to manipulate the shares they manage.

The non-technical dimension is also important. The vote verification process is extremely simple, accessible to any non-technical voter. This detail is critical for the population to accept the guarantees offered by this system, as opposed to others that are just as good but are extremely complex.

It should be mentioned that the threat model and the system assumption are realistic. On the contrary, other proposals are secure only under very lax criteria.

In summary, the proposal is considered to add value to the current state of electronic voting research. It is also flexible and easy to extend, so additions such as failure recovery protocols or as security layers against coercion could be easily added to this base.

Indeed, the two most immediate priorities for improving this proposal are:

- An exhaustive search for a solution to the coercion problem, to fully cover the list of desirable properties.
- Defining automatic protocols for real-time corruption detection and resolution, according to the cases we have mentioned previously in the analysis. The means for voters to report any error or manipulation are already provided, but the real-time resolution protocol is yet to be defined.

Once both requirements are met, this proposal satisfies all the current needs of e-voting systems. From this point, it is possible to think about developing an implementation, for a real future use. This would require a study on the platforms to be used, both on the server and client side, and on the scope of adoption.

Author Contributions: All authors have made relevant contributions to this work. The proposal of the problem was proposed by M.T.-R. and developed in numerous meetings involving D.O., J.J.-S., J.M.G.-G. and I.M.-M., who later led different parts of the work. Although all the authors actively participated in all tasks including writing and editing the manuscript, the theoretical part of the paper was leaded by M.T.-R. and D.O. All authors have read and agreed to the published version of the manuscript.

Funding: M.T.-R. is funded by a predoctoral grant from the University of Alcalá. M.T.-R., I.M.-M. and J.M.G.-G. are partially funded by Project PID2019-104855RB-I00/AEI/10.13039/501100011033 of the Spanish Ministry of Science and Innovation, by Project SBPLY/19/180501/000171 of the Junta de Comunidades de Castilla-La Mancha and FEDER and by Project UCeNet (CM/JIN/2019-031) of the Comunidad de Madrid and University of Alcalá. D.O. is partially supported by Project PID2019-104129GB-I00/AEI/10.13039/501100011033 of the Spanish Ministry of Science and Innovation, and by H2020-MSCA-RISE project 734922—CONNECT.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Gibson, J.P.; Krimmer, R.; Teague, V.; Pomares, J. A Review of E-Voting: The Past, Present and Future. *Ann. Telecommun.* **2016**, *71*, 279–286. [\[CrossRef\]](#)
- Madise, Ü.; Martens, T. *E-Voting in Estonia 2005. The First Practice of Country-Wide Binding Internet Voting in the World*; Gesellschaft für Informatik e.V.: Bregenz, Austria, 2006.
- McDaniel, P.; Blaze, M.; Vigna, G. EVEREST: Evaluation and validation of election-related equipment, standards and testing. In *Ohio Secretary of State's EVEREST Project Report*; Office of the Ohio Secretary of State: Columbus, OH, USA, 2007.
- Jones, D.; Simons, B. *Broken Ballots: Will Your Vote Count?* CSLI Publications: Stanford, CA, USA, 2012.
- Peacock, T.; Ryan, P.Y.A.; Schneider, S.; Xia, Z. Chapter 69—Verifiable Voting Systems. In *Computer and Information Security Handbook*, 2nd ed.; Vacca, J.R., Ed.; Morgan Kaufmann: Burlington, MA, USA, 2013; pp. 1103–1125. [\[CrossRef\]](#)
- Adida, B. Helios: Web-Based Open-Audit Voting. *USENIX Secur. Symp.* **2008**, *17*, 335–348.
- Yang, X.; Yi, X.; Kelarev, A.; Han, F.; Luo, J. A distributed networked system for secure publicly verifiable self-tallying online voting. *Inf. Sci.* **2021**, *543*, 125–142. [\[CrossRef\]](#)
- Rivest, R.L. The Threeballot Voting System. 2006. Available online: <https://dspace.mit.edu/handle/1721.1/96593> (accessed on 6 December 2021).
- Popoveniuc, S.; Hosp, B. An Introduction to Punchscan. In *IAVoSS Workshop on Trustworthy Elections (WOTE 2006)*; Robinson College United Kingdom: Cambridge, UK, 2006; pp. 28–30.
- Chaum, D.; Essex, A.; Carback, R.; Clark, J.; Popoveniuc, S.; Sherman, A.; Vora, P. Scantegrity: End-to-End Voter-Verifiable Optical-Scan Voting. *IEEE Secur. Priv.* **2008**, *6*, 40–46. [\[CrossRef\]](#)
- Chaum, D.L. Untraceable electronic mail, return addresses, and digital pseudonyms. *Commun. ACM* **1981**, *24*, 84–90. [\[CrossRef\]](#)
- Reed, M.G.; Syverson, P.F.; Goldschlag, D.M. Anonymous connections and onion routing. *IEEE J. Sel. Areas Commun.* **1998**, *16*, 482–494. [\[CrossRef\]](#)
- Querejeta-Azurmendi, I.; Arroyo Guardado, D.; Hernández-Ardieta, J.L.; Hernández Encinas, L. NetVote: A Strict-Coercion Resistance Re-Voting Based Internet Voting Scheme with Linear Filtering. *Mathematics* **2020**, *8*, 1618. [\[CrossRef\]](#)
- Islam, N.; Alam, K.M.R.; Tamura, S.; Morimoto, Y. A new e-voting scheme based on revised simplified verifiable re-encryption mixnet. In *Proceedings of the 2017 International Conference on Networking, Systems and Security (NSysS)*, Dhaka, Bangladesh, 5–8 January 2017; pp. 12–20.
- Chaum, D. Blind signature system. In *Advances in Cryptology*; Springer: Berlin/Heidelberg, Germany, 1984; p. 153.
- Kumar, M.; Katti, C.P.; Saxena, P.C. A secure anonymous e-voting system using identity-based blind signature scheme. In *Proceedings of the International Conference on Information Systems Security*, Mumbai, India, 16–20 December 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 29–49.
- Darwish, A.; El-Gendy, M.M. A new cryptographic voting verifiable scheme for e-voting system based on bit commitment and blind signature. *Int. J. Swarm. Intel. Evol. Comput.* **2017**, *6*, 2. [\[CrossRef\]](#)
- Rivest, R.L.; Adleman, L.; Dertouzos, M.L. On data banks and privacy homomorphisms. *Found. Secur. Comput.* **1978**, *4*, 169–180.
- Yang, X.; Yi, X.; Nepal, S.; Kelarev, A.; Han, F. A secure verifiable ranked choice online voting system based on homomorphic encryption. *IEEE Access* **2018**, *6*, 20506–20519. [\[CrossRef\]](#)
- Yang, X.; Yi, X.; Nepal, S.; Kelarev, A.; Han, F. Blockchain voting: Publicly verifiable online voting protocol without trusted tallying authorities. *Future Gener. Comput. Syst.* **2020**, *112*, 859–874. [\[CrossRef\]](#)
- Nakamoto, S. Bitcoin: A peer-to-peer electronic cash system. *Decentralized Bus. Rev.* **2008**, 21260.
- Pawlak, M.; Ponsizewska-Maraña, A. Trends in blockchain-based electronic voting systems. *Inf. Process. Manag.* **2021**, *58*, 102595. [\[CrossRef\]](#)
- Jafar, U.; Aziz, M.J.A.; Shukur, Z. Blockchain for Electronic Voting System—Review and Open Research Challenges. *Sensors* **2021**, *21*, 5874. [\[CrossRef\]](#) [\[PubMed\]](#)
- Nair, P.R.; Dorai, D.R. Evaluation of Performance and Security of Proof of Work and Proof of Stake using Blockchain. In *Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, Tirunelveli, India, 4–6 February 2021; pp. 279–283.
- Shamir, A. How to Share a Secret. *Commun. ACM* **1979**, *22*, 612–613. [\[CrossRef\]](#)
- Liu, Y.; Zhao, Q. E-Voting Scheme Using Secret Sharing and K-Anonymity. *World Wide Web* **2019**, *22*, 1657–1667. [\[CrossRef\]](#)
- Cascudo, I.; David, B. Albatross: Publicly attestable batched randomness based on secret sharing. In *Proceedings of the International Conference on the Theory and Application of Cryptology and Information Security*, Daejeon, Korea, 7–11 December 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 311–341.
- Gupta, K.D.; Rahman, M.L.; Dasgupta, D.; Poudyal, S. Shamir's Secret Sharing for Authentication without Reconstructing Password. In *Proceedings of the 2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 6–8 January 2020; pp. 0958–0963.
- Benaloh, J. Ballot Casting Assurance via Voter-Initiated Poll Station Auditing. *EVT* **2007**, *7*, 14.
- Neumann, S.; Feier, C.; Volkamer, M.; Koenig, R. Towards A Practical JCJ/Civitas Implementation. In *INFORMATIK 2013*; Horbach, M., Ed.; Jahrestagung der Gesellschaft für Informatik: Koblenz, Germany, 2013.
- Juels, A.; Catalano, D.; Jakobsson, M. Coercion-Resistant Electronic Elections. In *Towards Trustworthy Elections*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 37–63.

32. Küsters, R.; Liedtke, J.; Müller, J.; Rausch, D.; Vogt, A. Ordinos: A Verifiable Tally-Hiding E-Voting System. In Proceedings of the 2020 IEEE European Symposium on Security and Privacy (EuroS P), Genoa, Italy, 7–11 September 2020; pp. 216–235. [\[CrossRef\]](#)
33. Neji, W.; Blibech, K.; Rajeb, N.B. Incoercible fully-remote electronic voting protocol. In Proceedings of the International Conference on Networked Systems, Marrakech, Morocco, 17–19 May 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 355–369.
34. Ylonen, T.; Lonvick, C. *The Secure Shell (SSH) Authentication Protocol*; Technical Report; RFC 4252; IETF: Fremont, CA, USA, 2006.
35. Judson, T. *Abstract Algebra: Theory and Applications*; Virginia Commonwealth University Mathematics: Richmond, VA, USA, 2009.
36. Nguyen, V.L.; Lin, P.C.; Cheng, B.C.; Hwang, R.H.; Lin, Y.D. Security and privacy for 6G: A survey on prospective technologies and challenges. *IEEE Commun. Surv. Tutorials* **2021**, *23*, 2384–2428. [\[CrossRef\]](#)
37. Culnane, C.; Schneider, S. A Peered Bulletin Board for Robust Use in Verifiable Voting Systems. In Proceedings of the 2014 IEEE 27th Computer Security Foundations Symposium, Vienna, Austria, 19–22 July 2014; pp. 169–183. [\[CrossRef\]](#)

Conclusions

” See you, Space Cowboy

— Cowboy Bebop

In this PhD thesis, we applied graph theory to diverse real-world problems like wireless network and opinion analysis. Thanks to graph-based structures and tools like proximity graphs or graph-based distance measures, we have been successful on contributing several articles to the literature. In addition, we applied other mathematical tools like cryptography, Shamir’s Secret Sharing Scheme and finite fields to electronic voting.

First, in the field of Wi-Fi networks, we contributed to the state of the art in two ways. On one hand, publishing our IEEE 802.11 graph model [Gim+23], which is a fast and efficient model that can be used to explore all kinds of techniques and phenomena happening in Wi-Fi networks around the interaction between frequency allocation, access point distribution and physical layouts. Taking advantage of this new model, we found benefits on using channel bonding when 13 channels are available in Wi-Fi 4 [Gim+22].

At the same time, we have found promising results when trying to improve current strategies for frequency channel allocation, both in a centralized manner (using simulated annealing and related heuristics) and in a distributed fashion (using automated negotiation). The graph model is powerful but also flexible, since we have been able to do research about Wi-Fi 4, Wi-Fi 5 and Wi-Fi 6.

Second, we have used graph theory to propose alternative methods for opinion analysis. *SensoGraph* is a graph-based efficient alternative to statistical methods for opinion analysis [Ord+19]. It benefits from the possibility of obtaining large amounts of data from remote participants in an automated way, by using web technologies. To validate our contribution, we developed a web service and deployed a large scale analysis session, using the largest dataset so far with the projective mapping data collection technique [Ord+21]. For this dataset, we compared the original clustering mode of *SensoGraph*, Gabriel graph, with a new clustering mode based on

weighted distances. The results obtained are of good quality. As expected, the two duplicated control samples are identified as the most similar pair of products.

In another article, we suggested the use of freely linked graphs [Lah+22], which proved to be a more accurate and fine-grained method than its counterpart, free-sorting.

Finally, we made an innovative proposal for an e-voting system [Tej+21] using a directed acyclic graph, cyclic mixnets, and Shamir's Secret Sharing in finite fields. It offers a simple and friendly verification mechanism while remaining secure under a more conservative threat model than other proposals.

Overall, we provided yet another indication that graphs are a powerful structure for modelling relational data. Computer networks and similarity analysis are two perfect examples. With the help of graph theory, we have been successful on studying these two research lines and electronic voting, publishing several contributions to the state of the art.

On a personal level, this PhD thesis has been a journey for me. It required from lots of effort and dedication, but also allowed me to participate and contribute to cutting-edge contributions to the state of the art. I have been able to learn and progress in many different areas, meet new people and new places.

7.1 Conclusiones

En esta tesis doctoral hemos aplicado teoría de grafos a problemas prácticos como la optimización de redes inalámbricas y el análisis de opinión. Gracias a estructuras basadas en grafos y herramientas de la teoría de grafos como las métricas de distancias o los grafos de proximidad, hemos conseguido contribuir con varias publicaciones al estado del arte. Además, hemos aplicado otras herramientas matemáticas como la criptografía, el secreto compartido de Shamir y los cuerpos finitos al voto electrónico.

En primer lugar, en el área de las redes Wi-Fi, hemos contribuido al estado del arte de dos formas. Por una parte, publicando un modelo de redes IEEE 802.11 rápido y eficiente basado en grafos [Gim+23], que puede usarse para explorar toda clase de técnicas y fenómenos sobre las redes Wi-Fi, simulando asignaciones de canales de frecuencia y posiciones de los puntos de acceso y los clientes. Aprovechando este nuevo modelo, hemos encontrado una mejora de rendimiento al usar la técnica de *channel bonding* en Wi-Fi4 cuando hay al menos 13 canales disponibles. [Gim+22].

En esta misma línea, hemos encontrado resultados prometedores al intentar mejorar las estrategias actuales para la asignación de canales, tanto las centralizadas (basadas en heurísticas como *simulated annealing*) como las distribuidas (utilizando negociación automática). El modelo basado en grafos es flexible, por lo que nos ha permitido investigar sobre Wi-Fi 4, Wi-Fi 5 y Wi-Fi 6.

En segundo lugar, hemos usado teoría de grafos para proponer métodos alternativos para el análisis de opinión. *Sensograph* es una alternativa a los métodos tradicionales basados en estadística que utiliza grafos [Ord+19] para ofrecer una mejor eficiencia. Esta mejora nos permite introducir un mayor volumen de datos, habilitando la recogida automatizada de datos a través de tecnología web. Hemos desarrollado un servicio web para este propósito y desplegado una sesión de análisis a gran escala [Ord+21], utilizando el conjunto de datos más grande usando en la técnica de recogida de datos conocida como *projective mapping*. Con este conjunto de datos, hemos comparado el modo original de *SensoGraph*, basado en grafos de Gabriel, con un nuevo modo basado en aristas con pesos dependientes de la longitud. Los resultados obtenidos son de buena calidad, al observar que las dos muestras duplicadas se identifican como el par de productos más similar.

En otro artículo, proponemos un nuevo método de entrada basado en grafos en llamado *free-linking* [Lah+22]. Con este método, los participantes deben establecer un grafo de similaridad, definiendo la existencia de cada arista de forma individual. Los resultados muestran relaciones más precisas y de mayor resolución en comparación con su contrapartida, *free-sorting*.

Finalmente, proponemos un innovador sistema de e-voting [Tej+21] que se basa en un grafo dirigido cíclico como topología de red, en *mixnets* cíclicas y en el secreto compartido de Shamir en cuerpo finito. Ofrece un mecanismo de seguridad simple sin perder seguridad, y se basa en un modelo de amenazas más cauto que otras propuestas.

Como conclusión, esta tesis supone una prueba más de que los grafos son estructuras potentes para modelar datos relacionales, siendo las redes de computadores y el análisis de similaridad dos ejemplos perfectos. Con la ayuda de la teoría de grafos, hemos conseguido estudiar ambas líneas, además de investigar sobre voto electrónico, y publicar varias contribuciones al estado del arte.

A nivel personal, esta tesis doctoral ha sido un viaje para mí. Aunque ha requerido mucho esfuerzo y dedicación, me ha permitido participar en propuestas vanguardistas al estado del arte. También me ha permitido aprender en muchas áreas, progresar como persona y conocer nuevos sitios y personas.

7.2 Future Work

The work enumerated on this PhD thesis opened several further research lines. In the field of wireless networks, we are currently working on two directions. First, using the Wi-Fi graph model allows us to improve the current frequency channel allocation strategies. The current de-facto standard is LCCS (Least Congested Channel Search) [Ach06]. It is a distributed and uncoordinated algorithm, where each access point chooses the least congested channel unilaterally. This strategy is rather simple but it is suboptimal due to its uncoordinated nature, since access points cannot communicate in order to find a better distribution for most of them. This is why there are other techniques published which outperform LCCS [Mar+19]. The most important drawback of these techniques is the initial requirement to adjust the different parameters for each setting, depending on the type of network, Wi-Fi version, number of channels, number of clients and number of access points. Our latest work in this field aims to offer an automatic optimization method that adjusts automatically to the characteristics of the network using the adaptive properties of Thermodynamic Simulated Annealing [VLH03]. The preliminary results show a performance fairly similar to the manually-adjusted parametric version, which works out of the box without prior adjustment.

At a less advanced stage, we are working on distributed optimization using completely unmediated automated negotiation. This is a distributed approach, much more appropriate since frequency channel distribution is a distributed problem. In close collaboration with colleagues at the Interactive Intelligence Group at the Technical University of Delft in the Netherlands, where I made a research visit in 2022, we are working with SAOP and MOPaC [MJ22] in order to allow access points to participate in the negotiation over channel assignment. As a further research line, we are considering implementing a whole new protocol designed specifically for Wi-Fi networks.

The e-voting system, *DiverSEC*, can also be developed further. It can be improved in order to satisfy the only remaining property, coercion resistance, and it can also benefit from minor adjustments. But most importantly, it still remains to be implemented, tested and analyzed from a practical point of view.

Regarding *SensoGraph* and opinion analysis, we are about to finish a new software version of this method. It will allow us to test and compare new methods with even larger datasets from a web application. In particular, we want to implement a mix between projective mapping and free-linking, and study the correlation between the clustering obtained from Gabriel Graph and a manually established set of edges.

We are also currently working with incomplete block designs. The objective is to be able to study similarity in a large set of samples, since it can be difficult for participants to judge relationships between many vertices at the same time. To solve that, we try to study similarity in a large graph by correlating the data obtained from smaller subgraphs containing just some of the vertices.

In sum, there are still many different possibilities for further research building on top of the contributions of this PhD thesis. Hopefully, we can follow this trend and keep working on wireless networks, opinion analysis and electronic voting with the help of graph theory and the tools we have developed so far.

7.3 Trabajo futuro

El trabajos resumido en esta tesis doctoral abre varias líneas de investigación, a su vez. En el campo de las redes inalámbricas, trabajamos actualmente en dos direcciones. En primer lugar, el uso del modelo Wi-Fi basado en grafos nos permite mejorar las estrategias de asignación de canales de frecuencia. El estándar actual es LCCS (*Least Congested Channel Search*) [Ach06]. Se trata de un algoritmo distribuido y descoordinado, en el que cada punto de acceso escoge unilateralmente el canal menos congestionado. Esta estrategia es bastante sencilla pero es subóptima debido a su naturaleza descoordinada, ya que los puntos de acceso no pueden comunicarse para encontrar una distribución que sea mejor para la mayoría de ellos. Por ello, se han publicado otras técnicas que superan a LCCS [Mar+19]. El mayor inconveniente de estas técnicas es la necesidad inicial de ajustar los diferentes parámetros para cada configuración, dependiendo del tipo de red, versión Wi-Fi, número de canales, número de clientes y número de puntos de acceso. Nuestra última investigación en este campo pretende ofrecer un método de optimización automática que se ajuste automáticamente a las características de la red utilizando las propiedades adaptativas de la heurística *Thermodynamic Simulated Annealing* [VLH03]. Los resultados preliminares muestran un rendimiento bastante similar entre la optimización automática y la optimización paramétrica ajustada manualmente.

En una fase menos avanzada, estamos trabajando en la optimización distribuida utilizando negociación automatizada sin ningún intermediario. Se trata de un enfoque distribuido, dado que la distribución de canales de frecuencia es un problema distribuido. En colaboración con investigadores del Grupo de Inteligencia Interactiva de la Universidad Técnica de Delft (Países Bajos), donde realicé una estancia de

investigación en 2022, estamos trabajando con SAOP y MOPaC [MJ22] para permitir a los puntos de acceso negociar sobre el uso de canales de frecuencia. Como línea de investigación adicional, estamos considerando implementar un protocolo completamente nuevo diseñado específicamente para redes Wi-Fi.

El sistema de voto electrónico, *DiverSEC*, también puede seguir desarrollándose. En particular, falta ajustar el sistema para que cumpla con la resistencia a la coacción, o para hacer correcciones y mejoras menores. Pero sobre todo, el sistema todavía debe implementarse, probarse y analizarse en un caso práctico.

En cuanto a *SensoGraph* y el análisis de opiniones, estamos a punto de terminar una nueva versión de la aplicación web. Nos permitirá probar y comparar nuevos métodos con conjuntos de datos aún mayores procedentes de la recogida web. En concreto, queremos implementar una mezcla entre *projective mapping* y *free-linking*, para estudiar la correlación entre la matriz obtenida a partir de Gabriel Graph y el conjunto de aristas definido sin algoritmo de agrupamiento.

También estamos trabajando con diseños de bloques incompletos. El objetivo es poder estudiar la similaridad en un gran conjunto de muestras, ya que puede resultar difícil para los participantes juzgar las relaciones entre muchos vértices al mismo tiempo. Para solucionarlo, intentamos estudiar la similitud en un grafo grande correlacionando los datos obtenidos de subgrafos más pequeños que contienen sólo algunos de los vértices.

En resumen, todavía hay muchas posibilidades de seguir investigando a partir de las aportaciones de esta tesis doctoral. Esperemos que podamos seguir esta tendencia y seguir trabajando en redes inalámbricas, análisis de opinión y votación electrónica con la ayuda de la teoría de grafos y las herramientas que hemos desarrollado hasta ahora.

Bibliography

- [Ach06] Murali Achanta. “Method and apparatus for least congested channel scan for wireless access points”. US Patent App. 10/959,446. Apr. 2006 (cit. on pp. 88, 89).
- [Cha81] David L. Chaum. “Untraceable electronic mail, return addresses, and digital pseudonyms”. In: *Communications of the ACM* 24.2 (Feb. 1981), pp. 84–90 (cit. on p. 8).
- [Cos21] Michele Coscia. *The Atlas for the Aspiring Network Scientist*. arXiv:2101.00863 [physics]. Feb. 2021 (cit. on p. 1).
- [Del+34] Boris Delaunay et al. “Sur la sphere vide”. In: *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk* 7.793-800 (1934), pp. 1–2 (cit. on p. 2).
- [GS69] K. Ruben Gabriel and Robert R. Sokal. “A New Statistical Approach to Geographic Variation Analysis”. In: *Systematic Biology* 18.3 (Sept. 1969), pp. 259–278 (cit. on pp. 2, 5).
- [Gim+23] Jose Manuel Gimenez-Guzman, Ivan Marsa-Maestre, Luis Cruz-Piris, David Orden, and Marino Tejedor-Romero. “IEEE 802.11 graph models”. en. In: *Alexandria Engineering Journal* 66 (Mar. 2023), pp. 633–649 (cit. on pp. 3, 4, 11, 85, 86).
- [Gim+22] Jose Manuel Gimenez-Guzman, Ivan Marsa-Maestre, David Orden, Susel Fernandez, and Marino Tejedor-Romero. “On the Benefits of Channel Bonding in Dense, Decentralized Wi-Fi 4 Networks”. en. In: *Wireless Communications and Mobile Computing* 2022 (Feb. 2022). Publisher: Hindawi, e8497585 (cit. on pp. 3, 4, 29, 85, 86).
- [KR85] David G. Kirkpatrick and John D. Radke. “A Framework for Computational Morphology”. en. In: *Machine Intelligence and Pattern Recognition*. Ed. by Godfried T. Toussaint. Vol. 2. Computational Geometry. North-Holland, Jan. 1985, pp. 217–248 (cit. on p. 2).
- [LAH18] Jacob Lahne, Hervé Abdi, and Hildegard Heymann. “Rapid sensory profiles with DISTATIS and Barycentric Text Projection: An example with amari, bitter herbal liqueurs”. en. In: *Food Quality and Preference* 66 (June 2018), pp. 36–43 (cit. on p. 7).
- [Lah+22] Jacob Lahne, Katherine Phetxumphou, Marino Tejedor-Romero, and David Orden. “The free-linking task: A graph-inspired method for generating non-disjoint similarity data with food products”. en. In: *Food Quality and Preference* 95 (Jan. 2022), p. 104355 (cit. on pp. 3, 7, 86, 87).

- [Mar+19] Ivan Marsa-Maestre, Enrique de la Hoz, Jose Manuel Gimenez-Guzman, David Orden, and Mark Klein. “Nonlinear Negotiation Approaches for Complex-Network Optimization: A Study Inspired by Wi-Fi Channel Assignment”. en. In: *Group Decision and Negotiation* 28.1 (Feb. 2019), pp. 175–196 (cit. on pp. 88, 89).
- [MP13] Gary L Mullen and Daniel Panario. *Handbook of finite fields*. Vol. 17. CRC press Boca Raton, 2013 (cit. on pp. 2, 8).
- [MJ22] Pradeep K. Murukannaiah and Catholijn M. Jonker. *MOPaC: The Multiple Offers Protocol for Multilateral Negotiations with Partial Consensus*. arXiv:2205.06678 [cs]. May 2022 (cit. on pp. 88, 90).
- [Ord+19] David Orden, Encarnación Fernández-Fernández, José M. Rodríguez-Nogales, and Josefina Vila-Crespo. “Testing SensoGraph, a geometric approach for fast sensory evaluation”. en. In: *Food Quality and Preference* 72 (Mar. 2019), pp. 1–9 (cit. on pp. 5, 41, 85, 87).
- [Ord+21] David Orden, Encarnación Fernández-Fernández, Marino Tejedor-Romero, and Alejandra Martínez-Moraian. “Geometric and statistical techniques for projective mapping of chocolate chip cookies with a large number of consumers”. en. In: *Food Quality and Preference* 87 (Jan. 2021), p. 104068 (cit. on pp. 3, 5, 6, 41, 85, 87).
- [OT] David Orden and Marino Tejedor-Romero. “Registro Territorial de la Propiedad Intelectual de la Comunidad de Madrid número M-001652/2020” (cit. on p. 5).
- [Pag05] Jérôme Pagès. “Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley”. en. In: *Food Quality and Preference* 16.7 (Oct. 2005), pp. 642–649 (cit. on p. 5).
- [Sha79] Adi Shamir. “How to share a secret”. In: *Communications of the ACM* 22.11 (Nov. 1979), pp. 612–613 (cit. on p. 8).
- [TO] Marino Tejedor-Romero and David Orden. *SensoGraph Web* (cit. on p. 5).
- [Tej+21] Marino Tejedor-Romero, David Orden, Ivan Marsa-Maestre, Javier Junquera-Sanchez, and Jose Manuel Gimenez-Guzman. “Distributed Remote E-Voting System Based on Shamir’s Secret Sharing Scheme”. en. In: *Electronics* 10.24 (Jan. 2021). Number: 24 Publisher: Multidisciplinary Digital Publishing Institute, p. 3075 (cit. on pp. 3, 65, 86, 87).
- [VLH03] Juan de Vicente, Juan Lanchares, and Román Hermida. “Placement by thermodynamic simulated annealing”. en. In: *Physics Letters A* 317.5 (Oct. 2003), pp. 415–423 (cit. on pp. 88, 89).

List of Figures

1.1	Wi-Fi network of a residential building modelled in a graph, without interference edges. Extracted from [Gim+23]	4
1.2	Adjacency matrix which shows all edge weights in the 'distances' clustering mode.	6
1.3	Information flow of the vote shares in the cyclic graph topology	9

