MNRAS 000, 1–8 (2022) Preprint 16 March 2024 Compiled using MNRAS LATEX style file v3.0

Improving network selection and resource management through third-party monitors in heterogeneous networks

Zian Lin,[1] Cyrus Pellet[2]
[1] University of California – San Diego, San Diego, California, United States
[2] Ecole Polytechnique, Palaiseau, France

16 March 2024

ABSTRACT

As we enter the fifth-generation (5G) of cellular network technology, the increase in end users' expected Quality of Experience (QoE) and standards of privacy have given rise to challenges for network providers in providing a uniform and satisfactory user experience (UE) while minimising cost to the overall system as a priority. In order to accommodate this, new network selection and resource management (NSaRM) algorithms must be devised. To this end, we first conduct a survey of existing NSaRM algorithms along with their benefits and drawbacks, and provide common network scenarios in which such algorithms would be needed. We evaluate which approach against a set of qualitative criteria that describe sought-after characteristics of the network selection process. Following this, we propose a novel network selection model in which a third-party monitor delivers recommendations to users solely based on non-privileged information that any client would have access to.

Key words: Heterogeneous networks – Optimisation techniques – Network selection – Resource allocation

1 INTRODUCTION

Recent developments in wireless technologies have completely revolutionised the way we communicate and access information around the globe. A wide range of technologies have been continuously improving over the past decades to provide users with instant, high-throughput services of broadband access, with recent concerns such as mobility in mind. On one hand, standards of wireless local area networks (WLANs) such as IEEE 802.11a have been established and agreed upon by equipment manufacturers to provide high-speed data transmission services to users. (Wang & Kuo (2013)) On the other hand, wireless wide area networks (WWANs) have quickly evolved from the Global System for Mobile Communications (GSM) to Long Term Evolution (LTE), providing ubiquitous coverage and seamless mobility, even when roaming abroad. (Liang & Yu (2018)) On top of those, wireless personal area networks (WPANs) using technologies such as Bluetooth, Zigbee and LoRa, have been developed to bridge gaps in short-range coverage, enabling completely new areas of research to emerge in IoT and mobile computing. (Lei et al. (2013))

All of these networks have been progressively deployed with overlapping signal coverage, continuously bringing new uses and capabilities to users, eventually giving rise to what are now referred to as Heterogeneous Wireless Networks (HetNets). The coexistence situation of multiple networks and Radio Access Technologies (RATs) is critical to the operation of a wide range of devices with unique functionalities, and it can be anticipated that such deployments will only grow in the future. (Liu et al. (2014)) As such, the study of HetNets is critical to addressing the growing demand for data and the need for a more efficient use of the available radio spectrum.

HetNets differ from traditional wireless networks in several key ways. As previously mentioned, one of those is that HetNets integrate multiple types of network technologies, such as cellular networks and WLANs, whereas traditional networks are typically based on a single type of radio transmission technology. This allows HetNets to provide users with improved coverage, capacity, and performance in a multiplicity of scenarios.

Another key difference is that HetNets use a range of specialised technological components, such as Femto and Pico Access Points (FAP and PBS), coordinated multipoint (CoMP) transmission (Ali (2014)), and advanced antenna systems, to improve the efficiency of wireless spectrum use and support high-speed data rates. Traditional networks, on the other hand, typically rely on fewer, larger base stations and do not rely widely on these advanced technologies.

Moreover, HetNets are designed to support a wide range of User Equipments (UEs) and services, including smartphones, tablets, laptops, and IoT devices. Traditional networks, on the other hand, are typically designed to support a more limited range of devices and services with much more control over their operating patterns. Overall, HetNets represent a significant advance in wireless technology, offering improved performance and flexibility compared to traditional networks.

One of the commonly studied challenges of HetNets is the problem of efficient Resource Allocation (RA), which refers to the allocation of network resources such as spectrum and power to different UEs and services within the network. (AlSobhi & Aghvami (2019)) Indeed, one of the key issues in resource allocation for HetNets is the need to balance the conflicting goals of maximising the overall performance of the network, while also ensuring that individual users receive sufficient resources to meet their needs, ensuring an Always Best Connection (ABC). This can require complex algorithms and coordination mechanisms to ensure that resources are allocated in a fair and efficient manner.

The dynamic nature of HetNets can make it difficult to predict and manage the demand for resources. For example, the arrival of a large influx of users or the use of bandwidth-intensive applications (large file downloads for instance) can suddenly increase the demand for resources, requiring real-time adjustments to the allocation of these resources. (Odhiambo & Best (2013)) Amongst these allocation concerns, the problem of Network Selection (NS) is at the core of ensuring suitable resources are initially allocated for a client's specific use. The core principle of HetNets is that devices may have access to multiple RAT deployments, and the choice of which network to use can have a significant impact on the performance and reliability of the resulting connection.

This variety of choices in RATs allow users to be vertically handed off between different networks, but it is here that we encounter a dilemma. Delegating the network selection to the network providers may simplify the maximisation of QoE among users and minimise strain to the overall network, but the pursuit of optimality can take away user freedom in network choices. For example, a user streaming video content on their phone may want to stay on Wi-Fi in order not to incur data charges on their LTE plan, though they could inadvertently join the undesired network based on an imposed network selection. On the other end of the decision spectrum, delegating the network selection to the client can potentially result in high QoE for users, but such greedy selections may hamper overall network performance and potentially hurt others.

One example of this would be a scheme which selects a Wi-Fi AP for its higher throughput than the user's current LTE AP: nomadic users travelling by an AP for a public space (e.g. mall, library, square) would compete for capacity against the static users in the public space, and QoE would be degraded for people intending to use the Wi-Fi continuously. It is due to this dilemma between optimising for the network vs. optimising for the users that we must look into other Network Selection and Resource Management (NSaRM) schemes; ones that balance user freedom and overall network performance.

Considering such a tradeoff, as well as the assumption that a more optimal strategy can be devised without requiring radical changes to the network operators' deployments, we are therefore interested in answering the following question: **"Is there a non-greedy approach to network selection that does not rely on network providers' cooperation?"**. We hereby provide a survey of existing approaches of NSaRM algorithms, and propose a new solution involving the presence of a Third-Party agent to continuously monitor network conditions and issue recommendations to the end user. We discuss the mechanisms behind this approach, scenarios in which it can be beneficial, as well as ways to evaluate and quantify its improvement over other methods.

2 BACKGROUND

In order to conduct an informed survey of existing NSaRM methods, we are interested in collecting sensible and consistent definitions for the purposes of our study. The following help us describe and quantify the environment in which our survey takes place:

2.1 Heterogeneous Networks

We shall define HetNets as multi-radio networks comprised of hierarchical, multitier deployments of increasingly smaller cells, from MBS to small cells, including microcells, picocells, femtocells, and MBS remote access points such as remote radio heads (RRHs) and relays. Within these deployments, each User Equipment (UE) may employ multiple Radio Access Technologies (RATs) to communicate with network infrastructure.

Some literature restricts the study of HetNets to single carrier usage with spectrum cohabitation, however we shall here encompass distinct carrier deployments with multi spectral allocation. This allows us to take into account spectrum resource sharing between cells, and better represent typical modern scenarios such as malls, airports, etc... Nonetheless, even though base stations themselves are seen as self-organising and tightly coupled within the scope of a single operator, we shall assume that no cooperation of any kind is carried out from carrier to carrier.

Lastly, we recognise that HetNets rely on processes such as Admission Control/Cell Association/Power Control/Resource Allocation/Offloading decisions for the user to obtain optimised performance in their context. As stated before, we shall focus on Network Selection and Resource Management on the client side, though the steps listed above shall be acknowledged as dependent on the success of the device's Network Selection process.

2.2 User Equipment (UE)

In order to establish a connection over the local network or the internet through HetNet infrastructure, clients terminals comprise multiple radio access interfaces designed to send and receive data over multiple different channels across RATs.(Wang & Kuo (2013)) Following recent developments in enabling seamless mobility, one can identify two types of terminals:

- Multi-node terminals: fitted with multiple interfaces that do not support IP handover. Losing the connection on an interface necessarily results in a service interruption for the user. Moreover, individual sessions are limited to a unique interface during the whole operation life-cycle.
- Multi-homed terminals: have the ability to switch between networks without interrupting service, as well as share load for a single session across multiple simultaneous connections on multiple interfaces.

It is important to note that both multi-node and multi-homed terminals are critically reliant on appropriate network selection, anywhere, at any time, in order to maintain standards of an Always Best Connection (ABC). Whether this selection process is performed locally or offloaded externally, appropriate selection must consider QoS requirements for the user's activity, avoidance of congested networks, transient AP availability, and minimising the cost of unnecessary handovers, to cite a few factors.

2.3 User Experience

The coexistence of multiple RATs is key in the conceptualisation and deployment of HetNets. It entails that there is a pool of different wireless networks available for the UE to choose from in order to provide an optimal quality of experience to the user. There are many dimensions to the characterisation of user experience on a network, nonetheless, there exists no universally agreed upon notion of network quality. Moreover, while specifications of a network and a piece of UE are generally assumed to remain constant, many more metrics are highly dependent on its current configuration and current state.

User experience criteria can usually be classified into two different categories. The first one is referred to as Quality of Service (QoS) and encompasses the technical characteristics of a network connection. QoS signals include, but are not limited to:

- Signal strength (dBm)
- Throughput (Mbps) - Downlink or Uplink
- Packet loss (%)
- Available bandwidth (bps)
- Airtime and channel utilization (%)
- Bit error rate (%)
- Transmission rate (Hz)
- Latency (ms)
- Success & retry rates (%)
- Beacon availability (%)
- Jitter (UI)

In existing literature, QoS-focused schemes impose a set of constraints to maximise an aggregate network performance utility based on an indicator amongst those presented above, assuming that an increase or decrease in such indicators always results in an improvement in experience quality. Unfortunately this is not always the case,

The second category of criteria considers application-specific metrics and is referred to as Quality of Experience (QoE). Such metrics might include, depending on the scenario:

- Voice/Video Quality (MOS)
- Time to Interactive (s)
- Playback initial delay (s)
- Energy consumption (mAh)
- Peak Signal-To-Noise Ratio (dB)
- Structural Similarity Index (SSIM)

The QoE-focused schemes directly maximise the aggregate quality utility based on an indicator of the experience perceived by the user, yet also considering a set of constraints. Seufert et al. (2021) note that "By considering QoE aspects during the development process, it can be achieved that applications become network-aware by design". For instance, the advent of fifth generation broadband cellular technology (5G) unlocks a whole host of new uses for network connections, reinforcing the importance of tailoring infrastructure to the user. Hence, there is much to learn about the benefits of QoE-aware traffic management, and metrics such as described above should be taken into account to maximise network selection efficiency.

2.4 Traffic classes

Given the importance of QoE factors, described above, we understand that the user's activity on the network necessitates different requirements to consider when aiming to provide an optimal service. Looking at the common QoS signals listed previously, it is clear that latency plays a critical role in time-sensitive applications such as video streaming and gaming, but is much less of a concern when logging data from IoT devices, for instance. It is therefore necessary to establish a different set of UE factors to consider for each given traffic class the end user belongs to. We consider the following traffic classes, along with their associated relevant QoS and QoE signals ((aug)):

(i) Web Browsing
QoS: Downlink Bandwidth
QoE: Time To Interactive
Pattern: Large variable asymmetric downstream spikes

(ii) RTC Calls
QoS: Throughput, Jitter
QoE: Voice Quality, Peak Signal to Noise Ratio
Pattern: Medium constant symmetric rectangle bursts

(iii) Video Streaming
QoS: Downlink Bandwidth
QoE: Video Quality, Playback Initial Delay
Pattern: MTU-sized symmetric clustered rectangle bursts

(iv) Media Upload
QoS: Uplink Bandwidth, Packet Loss
QoE: Structural Similarity Index
Pattern: MTU-sized asymmetric upstream bursts

(v) Periodic Logging
QoS: Packet Loss
QoE: Energy Consumption
Pattern: Occasional small upstream dirac bursts

(vi) Latency sensitive
QoS: Latency, Throughput
QoE: Signal to Noise Ratio
Pattern: Constant rectangle bursts but variable

It is interesting to note the implications of the user's activity on network conditions. However, it should be noted that most NSaRM algorithms in the survey that follows disregard these distinctions in traffic types.

3 ASSESSMENT METHODOLOGY

In order to assess the characteristics of various network selection methods, we will take into account a set of qualitative criteria based on the objective of optimisation, simplicity and performance (Wang & Kuo (2013)). It is often the case that some network selection methods excel in certain situations and fall short in others due to assumptions and priorities set in their development: no NSaRM method is universally perfect. Regardless, we will attempt to describe the effectiveness of the surveyed methods against a set of qualitative factors as follows:

- **Simplicity**: amount of resources needed on client terminals to implement the method
- **Speed**: comparative speed of the decision process on cold start
- **Preference awareness**: scheme's consideration of user-oriented preferences (user's benefit), enabling compatibility for traffic classes
- **Network awareness**: scheme's consideration of traffic load (collective network's benefit)
- **Mobility**: scheme's degree of support for vertical client handover

4 SURVEY OF EXISTING APPROACHES

We shall consider the various schemes and mechanisms proposed in the available literature to analyse their various benefits and drawbacks and to assess whether such schemes could be of use in formulating our new network selection scheme. There are many network selection algorithms, ranging from single-attribute decision making (SADM), multiple-attribute decision making (MADM), game theory algorithms, and artificial neural network algorithms (ANN). The benefits and drawbacks of these techniques will be briefly covered in our survey.

Additionally, the literature also explores game theory schemes, artificial intelligence schemes that utilize machine learning and neural networks, Markov schemes, fuzzy logic schemes, utility theory schemes, etc. We will survey some of the most relevant approaches below.

4.1 Single-attribute decision making network selection algorithms

Gimenez et al. (2015) proposes two simple motives for optimal network selection, both based solely on the single attribute of throughput.

1. Always move a user to the cell with the highest estimated post-connection throughput. Can result in large amount of handovers, as this scheme does not consider the impact of the user handover on existing connections to the cell.

2. If and only if the estimated throughput of all active system users would increase upon handover, then move user to the cell with the highest estimated post-connection throughput. Reduces handovers, but forces throughput to remain at a 'local maxima' due to the fact that negative changes in estimated system throughput are discouraged.

These two solutions share similarities with the two hypothetical network selection schemes outlined in our introduction: either users are optimized for with greedy selections always connecting to the best cell, or the network is optimized for with total throughput being maximized. We can therefore find some insight into the benefits and drawbacks of optimizing for the users vs. optimizing for the network.

They find that optimizing for users tends to bring the best gain in throughput, albeit handover rates of up to three handovers per second can be reached at especially high loads. Optimizing for networks provides a lower gain in throughput, but the maximum handover rate can be cut by up to a whole order of magnitude. Comparing the final results for the two solutions, our first solution yields us a throughput gain of 107 percent for low-load conditions, 98 percent for medium-load conditions, and 90 percent for high-load conditions. In contrast, our second solution yields us a throughput gain of 69 percent for low-load conditions, 36 percent for medium-load conditions, and 18 percent for high-load conditions. From these results, we can deduce that if our solution is aiming to maximize throughput, it should lean more towards optimizing for the users rather than optimizing for the network.

The benefits of this scheme are that it's simple, high speed, and works fairly well in both low-mobility and high-mobility cases, with equal improvement in both scenarios. Due to the fact that throughput lowers with high user load, the network self-balances load and ensures that not too many users are connected to the same AP, making for a medium evaluation in network awareness. Unfortunately, it does not consider user traffic classes at all and the high handover rates can result in data interruptions for the users, resulting in a low evaluation for preference awareness.

Andreev et al. (2014) proposes three candidate solutions similar to Gimenez et. al's, one of which utilizies probabilistic selection in order to reduce handovers, a large problem with the aforementioned approach.

1. A user-centric approach that monitors neighboring APs and switches from LTE to Wi-Fi when a neighboring AP exceeds an SNR of 40 dB. This scheme is limited in dense urban environments due reasons similar to the example given in the introduction: nomadic users may increase load conditions for static users using Wi-Fi.

2. A RAN-assisted probabilistic approach that assigns a user to a RAT if it improves expected throughput and if a randomly generated number between 0 and 1 is above a numerical threshold determined by recent connections to the AP and reconnection probability. Reduces handovers and improves 'fairness', or the deviations between user throughput and average cell throughput.

3. Finally, a scheme in which the base station assigns users to networks based on the received signal strength indication (RSSI) it receives from the users. Similar to scheme 1, only the handover decision is made by the base station rather than the user device.

Their results found that scheme 2 resulted in the best individual throughput. This improvement was largely found solely at cell edges, with improvements in throughput reaching around 80 percent, compared to around 10 percent at best in cell centers. Such a scheme also leads to increases in energy efficiency (MB/J) and fairness (around a 50 percent increase in Jain's index). With this in mind and knowing the throughput statistics found in Gimenez et al. (2015), it seems that approach number 2 would be the most attractive option to explore further: we still connect the users to the most optimal cell much as we did with Gimenez's first solution, but we introduce hysteresis based on recent connections and reconnection probability in order to reduce high handover rates. With this, we have a slight decrease in simplicity and decrease in speed, but handovers are reduced, decreasing latency and the chances of data interruptions and improving preference awareness.

4.2 Multiple-attribute decision making network selection algorithms

As before mentioned, there are several common weighting methods when it comes to MADM algorithms: SAW, TOPSIS, MEW, GRA, ELECTRE, VIKOR, and WMC being some of the most prominent. Each weighting scheme has its own personal benefits and drawbacks. SAW, VIKOR, and TOPSIS tend to perform well for voice connections due to their lower jitter and packet delay values, while GRA and MEW tend to perform well for data connections due to their selection of networks with highest available bandwidths. SAW and TOPSIS are fairly fast and simple to implement, while ELECTRE and GRA are rather slow and complex. A bullet point list of characteristics can be found below, adapted from Martinez-Morales et al. (2010).

1. SAW (Simple Additive Weighting): Weights networks by calculating a weighted sum of multiple normalized parameters and ranking them. Fast and simple to implement. Good for voice connections. However, can select non-optimal networks due to the ability for highly positive parameters to outweight negative ones: e.g. picking a slightly cheaper network with low throughput as opposed to a slightly more expensive network with very high throughput.

2. TOPSIS (Technique for Order Preference by Similarity to Ideal Solution): Picks the network closest to the ideal solution and furthest from the worst existing solution. Like SAW, also fast and simple to implement and good with voice connections, but is sensitive to initial conditions like weights and user preferences, for better or for worse.

3. MEW (Multiplicative Exponential Weighting Method): Was created as a fix to problems with SAW. Uses weighted product instead of weighted sum. Diametrically opposed to TOPSIS in that it is not sensitive to initial conditions. Penalizes non-optimal networks heavier than SAW. Good with data connections.

4. GRA (Grey Relational Analysis): Creates a 'grey relational coefficient' depending on the magnitude of the highest upper bound in larger-the-better parameters, lowest lower bound in lower-the-better parameters, and moderate bound in nominal-the-best parameters. Ranks the networks in ascending order. Complex and lengthy, but handles many parameters and gives good performance for data connections.

5. ELECTRE (Elimination and Choice Expressing Reality): Weights attributes and pair-wise compares all networks to create a concordance set and discordance set. Picks the network with the highest average concordance index - discordance index. Complex and lengthy, but handles data connections well.

6. VIKOR (Viekriterijumsko Kompromisno Rangiranje/Multi-Criteria Optimization and Compromise Solution): Ranks networks based on a weighted average of the total difference between the highest weighted values given by a set of benefit parameters and the lowest weighted values given by a set of cost parameters. Handles many parameters well and is great for voice connections due to its resultant low packet delay and jitter at the cost of achieving a lower bandwidth than other options.

7. WMC (Weighted Markov Chains): Creates a Markov chain transition matrix where all matrix values are initialized by increasing the transition weight from network i to j for each decision factor q where i outperforms j in factor q. This increase is weighted by the importance of the decision factor. Distributes load well across networks and reduces number of vertical handoffs, though it doesn't quite reach the maximum bandwidth that other methods reach.

The above are examples of conventional MADM methods, though newer breakthroughs in MADM algorithms have combined weighting methods. Shi & Zhu (2012) tends towards a more complex algorithm: one that takes into account six objective network factors in addition to user preferences and traffic classes. The network factors of available bandwidth, peak data rate, packet delay, packet jitter, packet loss, and cost per bit are taken into account. Shi and Zhu's algorithm's complexity lies in the fact that they weight objective attributes (network factors) and subjective attributes (user preferences) in different ways: the former with entropy weighting (EW), the latter with an analytic hierarchy process (AHP). Once these attribute weight vectors are synthesized, they're integrated with group decision making (GDM) and then the consistency of the vector is analyzed, i.e. whether the objective and subjective decisions have been considered equally. Depending on the traffic class provided, the AHP decision matrices are weighted differently: higher weighting for packet delay and jitter for the conversational traffic class, higher weighting for packet loss and cost per bit for the interactive traffic class.

Utilizing the combination of these two weighting methods resulted in lower handoff numbers, lower packet delay, lower jitter, lower packet loss, and lowest cost per bit when compared to MADM methods that solely utilize either EW or AHP. This method is especially useful in cases where maximizing throughput is not the end goal: they provide real-time conversational traffic as an example of an instance where packet delay and jitter should be prioritized over total throughput. This method also takes into account user traffic classes, albeit only 'conversational' and 'interactive'. Overall, while this method is low in simplicity and low in speed, it is highly preference-aware and highly network-aware.

4.3 Game theory algorithms

Given that game theory is the study of decision-making, it stands to reason that it could be used as a tool to solve decisions on which network to connect to. This option has been regularly explored throughout the literature. Players, strategies, payoffs, and resources can be mapped to users/networks, available APs, QoS/QoE parameters, and resources like bandwidth and power. We attempt to tend towards a state of Pareto-optimal Nash equilibrium, a state where any other strategy combination would result in a decrease in payoffs.

Trestian et al. (2012) outlines at least 23 different approaches depending on whether the network decision process is classified as a user v. user game, a user v. network game, a network v. network game, and whether these games are cooperative or non-cooperative. While covering all possible combinations exceeds the scope of this paper, a notable example can be found in Antoniou et al. (2010).

Antoniou and et al. outline a cooperative user v. network repeated game where the payoff is defined by a utility function and the resource is bandwidth. Such a game is engineered to resolve in a solution that satisfies both user-satisfying and network-satisfying. It was found that the problem of user-network interaction is an analog of the famous Prisoner's Dilemma, and as such the user-network interactions can be treated as indefinitely iterated prisoner's dilemmas. They found that an adaptive-return strategy worked best, where 'trust' was built over continuous cooperations and non-cooperative behavior was punished by a user leaving the network for a number of periods proportional to the 'trust' that was built up.

Liu et al. (2014) on the other hand makes use of game theory as a bankruptcy game to calculate the Potential Contribution Ratio (PCR) of each network with the aim of establishing a ranking. The novelty here is to model each alternative network and attribute as a player and a kind of resource (or asset). Any subset of all alternatives is a coalition of players.

While game theory solutions like these provide deep insight into the benefits and drawbacks of user and network actions, it's difficult to manifest the theoretical payoff functions into functions of real parameters. Making decisions is a relatively fast process, resulting in high speed characteristics, but the complexity of implementing the abstract concepts results in a medium simplicity characteristic. The balance between preference awareness and network awareness depends on what variation of user v. user/user v. network/netwowrk v. network and cooperative/non-cooperative game you choose. Finally, such an approach as listed above would require the user and network to store each others history of interactions, using valuable storage space. Due to the fact that game theory algorithms require networks to have this 'memory' of an existing user/player, mobility support may prove difficult with the issue of maintaining trust parameters for nomadic users.

4.4 Artificial neural networks

With the past decade bringing large leaps in the field of artificial intelligence, it's clear that AI elements like neural networks could be applied to network selection algorithms.

Pahlavan et al. (2000) describes an ANN algorithm in which users send their received signal strengths (RSS) for all their discoverable networks to a third-party, middleware vertical handover manager. It can be thought of as an ANN twist on Andreev et al. (2014)'s single-attribute RSSI algorithm. The five most recent RSS samples of a given access point are fed into an ANN with one input layer, two middle layers, and an output layer. This neural network is trained to output either a 0 for no handover or a 1 for handover to the AP who's RSS samples were fed into the ANN. This is one of the earlier neural network approaches formulated for network selection, and as such it remains rather vague in terms of how the network is trained.

More recent research by Nasser et al. (2007) also describes an ANN algorithm which relegates decision making to a third-party, middleware vertical handover manager, albeit with a different internal structure. This manager consists of a network handling manager, a feature collector, and an ANN trainer/selector. The feature collector collects various user preference parameters like cost per minute, security, power consumption, or network conditions. These are fed as inputs to a backpropagation-based neural network consisting of an input layer, a hidden layer, and an output layer. With enough training, the network acts as a black box that selects the best network for the user based on the given parameters. Once a network has been selected as the best available option for the user, the network handling manager proceeds to connect the wireless device to the aforementioned network. In simulating their approach, they found that the optimal ANN had 5 input nodes, 10 hidden nodes, and 1 output node, achieving a performance rate over 99 percent.

The concept of using a third-party monitor to calculate the best network for the user is something that will be explored further on in the paper. The benefits of such a solution are obvious: all the calculations and externalities of network selection are offloaded to a third-party. Preference-awareness and network-awareness can be marked as high based on the parameters the ANN is trained on, and with the right parameters, mobility issues can also be solved with an ANN approach. With advances in neural networks, the speed of such a solution is likely medium-to-high. The disadvantages lie in the high complexity of such implementations and training the ANN models correctly.

	SADM	MADM	GT	ANN
Simplicity	High	Medium	Medium	Low
Speed	High	Medium	High	Medium
Preference awareness	Low	Medium	Low-High*	High
Network awareness	Medium	Medium	High-Low*	High
Mobility	Medium	Medium	Low	High

Table 1. Comparative table of surveyed algorithms. *(*Trades off with the other awareness parameter.)*

4.5 Summary

The survey above can be summarized simply in the table shown below.

It's important to note that the provided table is meant to be a comparative tool, not an objective evaluation, i.e. it aids in understanding that ANN algorithms are far more complex than SADM algorithms. For a detailed, complex evaluation, the survey above should be referenced. However, we can conclude that SADM algorithms tend to be simple and fast but have average performance in preference/network awareness and mobility. MADM algorithms have the most variation by far, with algorithms like SAW and TOPSIS being simple and fast while ELECTRE and VIKOR are complex and slow, but when it comes to awareness and mobility it improves on SADM in considering user preferences far better. Game theory algorithms largely suffer from concerns on how to implement them into real-world systems, resulting in medium simplicity and low mobility, but their preference and network awareness can be adjusted quite easily and they tend to be high speed. Finally, ANN algorithms are extremely good at considering preference awareness, network awareness, and mobility, but fall short in their medium speed and high complexity.

5 A NEW NSARM FORMULATION

Taking into the account the above NSaRM techniques, we look towards an algorithm that utilizes third-party monitors, much like the ANN system. This is because third-party monitors are able to strike a middle ground between the opposing factors of network optimality and user freedom while offloading computationally expensive calculations from both user and network. Additionally, network quality factors are highly variable both spatially and temporally as a function of the client distribution state: third-party monitors consider this and make optimal local decisions. Furthermore, such third-party monitors would only have a local overview of the network, and as such there would be no single operator with global visibility, further ameliorating privacy concerns.

In our proposition, users make a recommendation request to a third-party monitor. These third-party monitors identify six traffic classes from a user's network recommendation request as outlined in section 2.4: web browsing, RTC calls, video streaming, media upload, periodic logging, and latency sensitive traffic. These third-party monitors periodically ping nearby APs at regular intervals with requests of these traffic classes to evaluate their respective QoE factors also outlined in section 2.4 (e.g. making an RTC call request and evaluating throughput/jitter for QoS + voice quality/peak SNR for QoE). The recommendation exchange handshake takes on the structure in Appendix A of this document. In addition to this, the quality of service criteria is recorded continuously. Upon receiving back the QoE information from the preliminary ping, our third-party monitor creates a ranking for each usage class with APs that scored the highest on our QoE metric ranking near the top, and vice versa for the bottom. Hence, any users who are attempting to connect to a network while situated near the usage hotspot are able to be recommended the best APs available, all while only interacting with the third-party monitor.

In order to reduce the impact of its deployment to a minimum, we shall enable its operation with no additional data than what a normal client has access to, the only differences being its replication across a discrete set of points in the operation space, and its continuous operaation over time. These characteristics allows the system to have access to an overview of the HetNet's condition across space and time; the lack of which being one of the main limitations of a single isolated client. This approach hence aims to not only avoid the most sub-optimal distributions induced by greedy client-centric approaches, but also eliminate the coordinated actors assumptions of centralised approaches.

6 EVALUATION

We evaluate our method with several assumptions. We study a closed, static environment (in which mobility is not a concern), where two operators maintaining different RATs provide services to a set of UEs. These third party nodes are deployed across the area of interest, allowing us to survey the entire user space. Optimizing the placement of these nodes is a problem that can be studied further, but we will assume that the location of heavily trafficked areas are known a priori and can form the basis of the monitors' set locations. We also assume that recommendations are taken by the users.

Qualitatively, we can think of our approach as a middle ground between the more user-preference oriented MADM methods and the ANN third-party monitor methods. Despite our method considering six traffic classes, the baseline mechanism is of relatively high simplicity. A user simply sends a traffic class request to a third-party monitor, which has already prior calculated a ranking of the best networks to service each traffic class by sending dummy requests to APs and evaluating them based on their respectively associated QoE metrics. For this reason, our approach is high simplicity. With the ranking being pre-calculated by the third-party monitor at regular intervals, the response time is also high speed. With the network ranking mechanism being customized for each provided traffic class, it is also highly preference aware when compared to previous schemes that utilize universal QoE metrics for all traffic classes.

However, our formulation is not without its drawbacks. While it's likely to work with static and slow-moving nomadic users, the pre-calculation of the network rankings could lead to slight drops in high-speed mobility situations, due to its inability to adapt to rapidly changing network conditions. The optimal time period for re-calculating the network rankings is something that must be further confirmed with quantitative research. Additionally, because the third-party monitors only have make the optimal local decisions, it is not guaranteed that these optimal local decisions will emerge as optimal global decisions. Due to the above reasons, we'd comparatively characterize our formulation as medium in both network awareness and mobility.

The comparative evaluation of our third-party monitor (TPM) method with the current formulations in the literature can be found in the updated table above.

	SADM	MADM	GT	ANN	TPM
Simplicity	High	Medium	Medium	Low	High
Speed	High	Medium	High	Medium	High
Preference awareness	Low	Medium	Low-High*	High	High
Network awareness	Medium	Medium	High-Low*	High	Medium
Mobility	Medium	Medium	Low	High	Medium

Table 2. Comparative table of surveyed algorithms. *(*Trades off with the other awareness parameter.)*

7 CONCLUSIONS AND FURTHER WORK

In summary, this paper seeks to define a new network selection and resource management schema for heterogeneous networks. We first surveyed several modern NSaRM schemes and compared their various benefits and drawbacks with a five-factor evaluation criteria consisting of simplicity, speed, preference-awareness, network-awareness, and mobility. Then, we propose a new schema involving third-party monitors. Our schema utilizes many third-party monitors deployed over an active hotspot area that periodically issue recommendations to nearby users depending on their location and traffic class. We chose to enumerate six different traffic classes: web browsing, real-time calls, video streaming, media uploads, periodic logging, and latency-sensitive traffic.

The main advantage of our approach lies in our third-party monitors 'caching' the performance metrics for all possible traffic classes, meaning less computation for the user, less information provided between the user and the network, high speed, and high simplicity. Additionally, high granularity when it comes to traffic classes and the use of different QoE metrics depending on each traffic classes results in better preference-awareness. However, the 'caching' of network rankings could potentially lead to performance drops in high-speed mobility cases, and the practice of third-party monitors making optimal local decisions may not necessarily translate to optimal global decisions. Such drawbacks would have to be evaluated objectively through further network simulations.

Next steps for future research should involve further investigation towards optimizing the placement of our third-party monitors, traffic class selection, and evaluating our formulation's overall network performance and mobility performance with extensive simulations. The topic of placing the third-party monitors is a subject that was glossed over in this paper: it is to be determined whether elementary arrangements such as hexagonal packing are optimal for covering network hotspots. Additionally, it should be considered as to whether our current choices of traffic classes are optimal: future developments in technology may open up new avenues of traffic not covered with the purview of our six traffic classes. Finally, it is necessary to conduct more realistic network simulations involving multiple NSaRM schemes in order to evaluate their benefits and drawbacks under a noisy, real-world environments.

ACKNOWLEDGEMENTS

We'd like to extend our warm gratitude to Dr. Juan-Antonio Cordero Fuertes for his advice and guidance in the writing of this report: it could not have been written without his extensive networks expertise.

REFERENCES

AlSobhi W., Aghvami A. H., 2019, in 2019 26th International Conference on Telecommunications (ICT). pp 330–334, doi:10.1109/ICT.2019.8798829

Ali M. S., 2014, International Journal of Future Generation Communication and Networking, 7, 91

Andreev S., Gerasimenko M., Galinina O., Koucheryavy Y., Himayat N., Yeh S.-P., Talwar S., 2014, IEEE Wireless Communications, 21, 86

Antoniou J., Papadopoulou V., Vassiliou V., Pitsillides A., 2010, Computer Networks, 54, 2239

Gimenez L. C., Kovacs I. Z., Wigard J., Pedersen K. I., 2015,] 10.1109/VTC-Fall.2015.7391176, pp 1–5

Lei L., Zhong Z., Zheng K., Chen J., Meng H., 2013, IEEE Wireless Communications, 20, 34

Liang G., Yu H., 2018, EURASIP Journal on Wireless Communications and Networking, 2018, 241

Liu B., Tian H., Wang B., Fan B., 2014, in 2014 IEEE 11th Consumer Communications and Networking Conference (CCNC). pp 501–506, doi:10.1109/CCNC.2014.6866617

Martinez-Morales J. D., Pineda-Rico U., Stevens-Navarro E., 2010, pp 309–314

Nasser N., Guizani S., Al-Masri E., 2007. pp 5671 – 5676, doi:10.1109/ICC.2007.940

Odhiambo M., Best M., 2013.

Pahlavan K., Krishnamurthy P., Hatami A., Ylianttila M., Mäkelä J.-P., Pichna R., Vallstron J., 2000, Personal Communications, IEEE, 7, 34

Seufert A., Schröder S., Seufert M., 2021, SN Computer Science, 2, 463

Shi Z., Zhu Q., 2012, The Journal of China Universities of Posts and Telecommunications, 19, 92–98, 114

Trestian R., Ormond O., Muntean G.-M., 2012, IEEE Communications surveys & tutorials, 14, 1212

Wang L., Kuo G.-S. G., 2013, IEEE Communications Surveys & Tutorials, 15, 271

On Traffic Patterns of HTTP Applications, `https://hal.archives-ouvertes.fr/hal-00685658/document`

APPENDIX A: THIRD PARTY INTERFACE DEFINITION

```
enum TrafficClass
{
  T_WEB_BROWSING = 0;
  T_RTC_CALLS = 1;
  T_VIDEO_STREAMING = 2;
  T_MEDIA_UPLOAD = 3;
  T_PERIODIC_LOGGING = 4;
  T_LATENCY_SENSITIVE = 5;
}

message gHetNetRecoReq
{
    TrafficClass trafficClass = 1;
    map<string, int32> rSSI_samples = 2;
    ...
}

message gHetNetRecoRes
{
    repeated string gHetNetCellIDs = 1;
    ...
}
```

APPENDIX B: LIST OF ACRONYMS

3GPP: 3rd Generation Partnership Project
ABC: Always Best Connection
ANDSF: Access Network Discovery and Selection Function
AP/BS: Access Point/Base station
CC: Client Cooperation
CSZ: Channel State Information
CoMP: Coordinated Multipoint Transmission
DAS: Distributed Antenna System
DSL: Digital Subscriber Line
DoC: Duration of Connection
FAP: Femto Access Point
FFR: Fractional Frequency Reuse
FFZ: Femto-Free Zone
IM: Interference Mitigation
ISD: Inter-Site Distance
IoT: Internet of Things
LTE: Long Term Evolution
LoS: Line of Sight
MADM: Mutliple Attribute Decision Making
MBS: Macro Base Station
MOE: Macro cell User Equipment
MRC: Maximum Ratio Combining
MRT: Maximum Ratio Transmission
MTU: Maximum Transmission Unit
NSaRM: Network Selection and Resource Management
OFDMA: Orthogonal Frequency Division Multiple Access System
PAN: Personal Access Network
PBS: Pico Base Station
QoS: Quality of Service
QoE: Quality of Experience
RAI: Radio Access Technology
RAN: Radio Access Network
RRM: Radio Resource Management
RS: Relay Station
SE: Spectral Efficiency
SMD: Smart Mobile Device
SNR: Signal to Noise Ratio
UE: User Equipment
eNB: Evolved Node B

This paper has been typeset from a TEX/LATEX file prepared by the author.