Throughput Enhancement via Multi-Armed Bandit in Heterogeneous 5G Networks

Ankur Vora, Kyoung-Don Kang
State University of New York at Binghamton, NY, USA.
{avora4, kang}@binghamton.edu

***Abstract*— In heterogeneous networks, a user equipment (UE) can directly communicate with the macro base station (BS) or a small, low-power pico or femto BS. Alternatively, it can indirectly communicate with the macro BS through one or more intermediate device (UE) or a relay-station that uses the over-the-air backhaul to the macro BS. Due to the highly dynamic and uncertain nature of wireless communication, it is essential for a UE to choose an optimal communication mode and a neighbor to which it connects, e.g., a macro/small BS in the direct communication mode or a nearby relay/device in the indirect communication mode. In this paper, we apply an effective reinforcement learning method, called multi-armed bandit (MAB), to shed light on this problem. Especially, we apply MAB supported by the Thompson sampling theorem to pick an optimal arm—a neighbor that determines the communication mode and resulting performance, while effectively dealing with the exploration-exploitation dilemma in MAB. In a simulation study undertaken in Matlab, we compare the performance of the proposed approach to several baselines representing the current state of the art. Our approach enhances the throughput normalized to the optimal throughput by approximately 8–97% compared to several baselines representing the state of the art. Further, it improves the throughput by up to 15% compared to the best performing baseline [1], [2].**

I. Introduction

A major challenge faced by 5G network operators is to provide uniform coverage in metropolitan areas, which are densely filled with high-rise buildings and other objects. One way to expand the existing network is by adding more macro base stations (BSs); however, installing more macro BS sites incurs a significant capital expense for the network operator. To overcome this challenge, researchers have investigated new 5G networks based on small cells and low-power pico or femto BSs, and relays [3], [4], [5], [6]. A UE can directly communicate with a micro, pico or femto BS, which consumes less power than a macro BS does. A relay station is connected to a macro BS using the over-the-air backhaul to relay messages between the BS and UEs.[1] Further, device-to-device (D2D) communication is an important technology in heterogeneous networks. In sum, when leveraged effectively, small BSs, relay stations and D2D communication can enhance the network coverage, user density and service quality such as throughput in heterogeneous networks [7], [8].

According to a recent survey [9], a service provider in North America plans to install approximately 850,000 small cells by the end of 2025 out of which 400,000 are expected to be deployed by the end of 2018. In such a dense heterogeneous network, it is essential to make all small cells efficiently work together with the macro BS to enhance the network performance [10], [11]. One of the key challenges for a user equipment (UE) is to which BS, relay or another UE it should connect to get optimal network performance. In a heterogeneous network, a UE can be located at the edge of multiple cells or in a place where it can either directly communicate with a macro/small BS or indirectly communicate with the macro BS through a relay station or device (resulting in multihop communication). If a UE can predict connecting to which node, i.e., another UE, a relay, small BS, or the macro BS, is expected to be optimal, it can achieve higher throughput. However, optimizing throughput in a heterogeneous 5G network is very challenging due to the complexity of the network and the highly uncertain nature of wireless communication. To address this challenge, we apply reinforcement learning to optimize the throughput for downlink communication in a heterogeneous 5G network.

Reinforcement learning is a branch of machine learning that aims to maximize cumulative rewards (or minimize regrets due to suboptimal performance) without requiring prior training [12]. A challenge for reinforcement learning is the dilemma between exploitation of the current knowledge and exploration of the unexplored territories. The former aims to maximize rewards based on the available information, while the latter intends to gather more information. Exploring new choices or unexplored alternatives can be risky, potentially incurring regrets. On the other hand, always exploiting current knowledge may lead to a local optimum with suboptimal performance, failing to find a global optimum. To deal with the exploration-exploitation dilemma, we apply the well-established multi-armed bandit (MAB) methodology, which models a gambler that plays a row of slot machines to maximize the gain based on the probability distribution of the reward provided by each machine [12], [13].

In the context of a heterogeneous network considered in this paper, each UE (gambler) applies the MAB technique to maximize the reward, i.e., throughput, by carefully choosing a nearby UE, relay, small BS or macro BS that consist competing choices, i.e., arms/levers, for downlink communication when only limited information about the choices, e.g., their signal-to-interference-plus-noise ratio (SINR), channel state information (CSI), and channel bandwidth, is available [1].

[1]From a UEs perspective, a relay node acts as if it is a BS. On the other hand, from the perspective of a macro BS, relay nodes can be viewed as UEs.

A UE chooses direct communication with the macro, pico or femto BS or indirect communication through a relay or another UE via device-to-device (D2D) communication. A choice then returns a random reward based on its actual probability distribution unknown to the original UE a priori. Based on the actual reward, the UE updates the corresponding information, i.e., probabilistic distribution of rewards/regrets, of the neighbors (arms) and utilizes it to make decisions for optimal path selection in the future. In this paper, a UE applies MAB supported by the Thompson sampling theorem [13] for optimal arm selection. MAB with Thompson sampling can effectively deal with the exploration-exploitation dilemma in MAB, since it keeps updating the beta distribution based on the observed rewards and regrets of each arm and takes random samples from the distribution for optimal arm selection [13], [14]. Despite the effectiveness, related work on applying MAB with Thompson sampling to optimize the performance in a 5G heterogeneous network is relatively scarce.

In our simulation study, the performance of the proposed methodology is compared to the performance of several state-of-the-art MAB methods using different arm selection techniques: 1) the greedy [15], 2) Levy [16], 3) ϵ-greedy [17], 4) Boltzmann [18] and 5) upper confidence bound (UCB) method [1], [2]. The performance evaluation results show that our approach enhances the (normalized) throughput by 8–97% compared to the tested baselines. Further, it improves the throughput by up to 15% compared to the best performing baseline based on the UCB method [1], [2].

The rest of this paper is organized as follows. Section II discusses the overall network structure. Section III discusses the proposed methodology and the baselines. Section IV presents the performance evaluation results. Finally, the paper is concluded in Section V.

II. Heterogeneous Network Architecture and Communication Modes

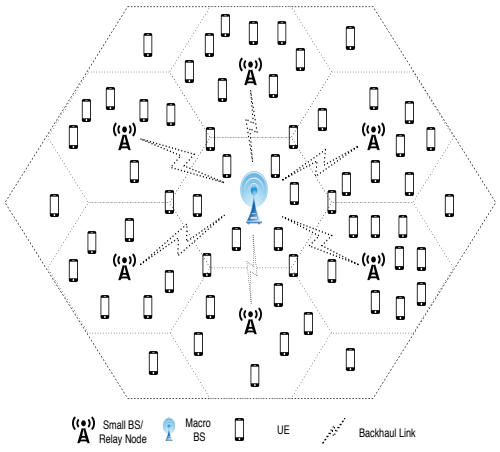

Fig. 1: Heterogeneous Network Scenario

A. Communication Scenario

Figure 1 shows an example structure of a heterogeneous network, in which a macro BS expands its coverage and overall performance by adding multiple small cells into one macro cell. It consists of one macro BS at the core of the cell connected with multiple small cells. These small cells may be pico, nano, femto cells or relay stations connected with the macro BS. A UE in these cells can connect to its nearest BS depending on its location in this network. A throughput achieved by a UE at time t depends on the path provided by a neighboring node, which is another UE, a relay, a small BS or the macro BS. In a heterogeneous network, communication takes place in one of two modes as follows:

1) Direct Communication Mode: If a UE communicates to its nearest macro or small BS directly, it is considered direct communication. Initially, a UE tries to connect to its nearest available BS and establish its presence over the radio access network (RAN). The UE connects to the BS from which it gets the higher SINR and CSI. In a case where the UE is at multiple cell edges, where it can receive data from multiple BSs, it applies the same strategy.

2) Indirect Communication Mode: If a UE communicates with macro BS through one or more device or relay, we consider it indirect communication. A UE can initiate communication by making a peer discovery of neighbors and establishing a connection to the macro BS as shown in Figure 2. In this process, the UE discovers other UEs in the proximity and exchanges control signals among them. Based on the received response, the UE decides the path based on their SINR and CSI. If a UE is outside the coverage as shown at the right bottom in Figure 1, it can discover another UE, which has a connection to the nearest BS and establishes a connection to a BS. In case of multihop communication, one hop may have a favorable channel condition whereas others may suffer from poor channel conditions. The overall throughput achieved over multiple hops is the minimum throughput of all the hops. Figure 2 depicts the indirect communication mode. For example, if UE A in Figure 2 directly communicates with the macro BS at the top of the figure, it can achieve 2Mbps.[2] However, if every UE performs a peer discovery and supports multihop communication, the UE can achieve 3Mbps, if it follows the path A→small-BS→macro-BS as shown in Figure 2. Hence, a UE needs to select the optimal path that provides the optimal end-to-end throughput [19]. Although the latency is incurred for the first communication with the small-BS, the following communications with the macro-BS through the small-BS is processed in a pipelined fashion. Thus, the impact is negligible. Further, the service coverage can be enhanced via indirect communication. For example, in Figure 2, UE E that is outside the macro cell can use the path E→D→macro-BS with 3Mbps throughput.

[2]The number on each dotted link in Figure 2 is the provided throughput in Mbps.

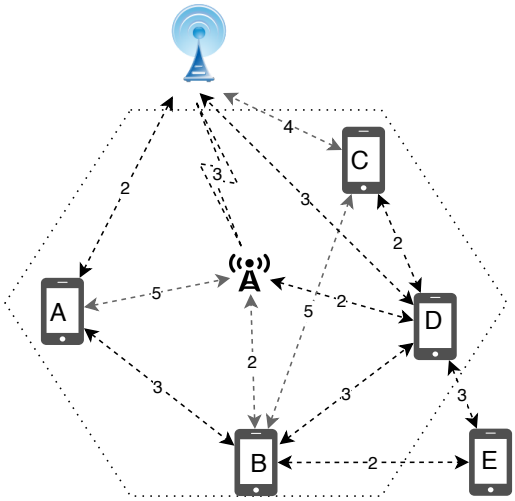

Fig. 2: Optimal Path Selection in Indirect Mode

III. Multi-Armed Bandit via Thompson Sampling

Ideally, a UE (player) pulls an arm of the bandit, i.e., selects the communication mode and a neighbor, e.g., the macro-BS, a small-BS, a relay or a UE, for wireless communication, which has *an optimal path with the highest probability to maximize the reward*, e.g., SINR or throughput, between the original UE and the macro-BS. However, achieving this objective is very hard in reality, since the actual probability is unknown a priori and wireless communication is highly dynamic and uncertain. To address the challenge, we apply the Thompson sampling theorem [13], [14] to deal with the exploration-exploitation dilemma in MAB. Although it is based on simple probability matching, Thompson sampling is known to provide similar performance to state-of-the-art approaches. Moreover, it significantly outperforms other alternatives such as the upper confidence bound (UCB), which is a popular method with strong theoretic guarantees on regrets, in some cases [20]. In this section, we introduce Thompson sampling to deal with the exploration-exploitation dilemma in MAB in comparison to the greedy method, which is one of the simplest methods for MAB. Further, we describe how to extend Thompson sampling to support communication mode selection in 5G. In Section IV, our approach substantially outperforms the state-of-the-art baselines including the greedy and UCB algorithms.

In MAB, there is a set of K actions (arms). In each round t, the player selects an arm a_t and observes the reward r_t for the chosen arm. There is a fundamental tradeoff between getting new information about rewards, i.e., exploration, and making optimal decisions using the available information, i.e., exploitation. For example, in a heterogeneous network, a UE can choose either a direct or indirect communication with the macro-BS. Also, when it chooses an indirect communication, there can be more than one peer, e.g., a relay, small-BS, or UE, that has a communication path to the macro-BS with potentially different end-to-end throughput. If one arm (e.g., a communication path) is always chosen, it is impossible to know if another arm is better (e.g., provides higher throughput). On the other hand, optimal performance (e.g., throughput) cannot be achieved if always different arms are chosen regardless of their relative performance.

In the Bernoulli bandit, an action that pulls an arm k produces a reward of 1 and 0 with probability θ_k and $1-\theta_k$, respectively [21], [20], [12]. Thus, θ_k is the action's success probability or mean reward where the mean rewards $\theta = (\theta_1, \ldots, \theta_K)$ are unknown in advance. In the first round, action a_1 is taken and reward $r_1 \in \{0, 1\}$ is produced with the success probability $P(r_1 = 1|a_1, \theta)$. The reward r_1 is observed and another action a_2 is taken, and the process is repeated. Based on the beta distribution with two shape parameters, α and β, the mean reward of arm k, e.g., path k, is estimated as:

$$\mu_k = \frac{\alpha_k}{\alpha_k + \beta_k} \tag{1}$$

In the Bernoulli bandit, α_k is incremented by 1 if pulling arm k produces a reward of 1. Otherwise, β_k is incremented.

In the **greedy** method for MAB, simply the arm with the $argmax_k \mu_k$ is pulled next. Thus, the greedy method essentially stops exploration once an optimal arm, which may turn out suboptimal later, is found. In **Thompson sampling** adopted in this paper, however, a *random sample* $\hat{\theta}_k$ is taken for each arm k based on the probability density function of the beta distribution for $0 \leq x \leq 1$:

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{2}$$

where Γ is the gamma function. Thus, the arm (path) with $argmax_k \hat{\theta}_k$ is chosen next to strike a balance between exploration and exploitation [21], [20], [12]. In the following, we describe how to extend the Bernoulli MAB for communication mode selection in 5G.

In this paper, a UE initializes $\alpha_k = \beta_k = 1$ for each neighbor k that has a direct or indirect path to the macro-BS if the UE has no prior distribution about its peers when, for example, it has just moved into the current cellular area. Thus, a neighbor is selected in a uniform random manner in such a case. For efficient communication mode selection, we propose to leverage two downlink synchronization signals in LTE and 5G—the primary synchronization signals (PSS) and secondary synchronization signals (SSS) in the physical layer—used by UEs to acquire the cell identity and frame timing.[3] We assume that each BS or any other node that can act as an intermediary between a UE and BS periodically disseminates PSS and SSS.[4] As PSS and SSS are required for cell identification and frame timing, leveraging them for communication mode selection incurs little extra overhead.

After a UE observes the average SINR of $N > 1$ PSS and SSS provided by a neighbor i, it updates the beta distribution parameters for the neighbor. In this paper, the time interval for N PSS and SSS, e.g., 10ms, is one measurement period.

[3]Our approach is not tied to PSS and SSS though. For example, heartbeat messages exchanged between devices can be used instead of PSS and SSS.

[4]In LTE, two PSS and SSS are broadcast every 10ms.

In the j^{th} period ($j \geq 1$), we consider the mean end-to-end SINR of the path from the UE to the macro-BS provided by its neighbor i as the estimated reward provided by the neighbor and use $\alpha_i(j)$ to denote it.[5] Further, we estimate β_i of the UE's neighbor i in the j^{th} period as follows:

$$\beta_i(j) = \begin{cases} \beta_i(j-1) & \text{if } \alpha_i(j) > \alpha_i(j-1) \\ \alpha_i(j-1) - \alpha_i(j) & \text{otherwise} \end{cases} \quad (3)$$

In this paper, the greedy algorithm used as a baseline in Section IV updates the mean reward of the UE's peer i based on $\alpha_i(j)$ and $\beta_i(j)$:

$$\mu_i(j) = \frac{\alpha_i(j)}{\alpha_i(j) + \beta_i(j)} \quad (4)$$

and selects the peer with $argmax_i\ \mu_i(j)$ next.

In Thompson Sampling, the probability density function in Eq 2 is updated using $\alpha_i(j)$ and $\beta_i(j)$ for each peer i of the UE. The UE then takes random samples from the probability density functions of its neighbors and selects the neighbor with the highest random sample value among all its neighbors to efficiently strike a balance between exploration and exploitation.

IV. Performance Evaluation

In Matlab, the performance of the proposed framework is evaluated using the LTE and statistics and machine learning toolboxes. All the wireless channels considered here are Rayleigh fading channels in nature. The UE in a network considered here supports the quadrature phase shift keying (QPSK) modulation scheme with 2×2 multiple-input multiple-output (MIMO) antennas. We are considering 1000 UEs in each cell. A UE in a cell has connectivity with one or more device, small cell, relay station and macro BS. A UE outside the cell area can do a neighbor discovery with a UE in a cell and establish a D2D communication.

In addition to the greedy algorithm discussed in Section III and used in [15], we consider the following state-of-the-art baselines for thorough performance comparisons:

- ϵ-greedy used in [17] is an improvement of greedy. It applies the greedy algorithm with probability $1 - \epsilon$, while choosing a random arm with probability ϵ to keep exploring and reduce regrets.
- Boltzman applied to wireless communication [18] is a heuristic that improves upon greedy. In Boltzman, an arm with the higher mean reward is chosen with a higher probability.
- A two-armed Levy bandit applied to device-to-device communication in [16] is used as a baseline.
- UCB used in [1], [2] is another baseline. At round t, UCB selects arm a_k with $argmax_{k \in K} \hat{Q}_t(k) + \hat{U}_k(a)$ among K arms where $\hat{Q}_t(k)$ is the estimated mean reward and $\hat{U}_k(a)$ is the estimated upper confidence of a_k.

[5]The end-to-end SINR of a path is determined by the minimum SINR of the link(s) on the path.

In this section, we observe that the proposed approach based on Thompson sampling significantly outperforms the baselines, because greedy stops exploration once it finds an optimal arm based on the current knowledge and ϵ-greedy and Boltzmann are heuristics that extend greedy. As a two-armed Levy bandit considers only two arms (a safe arm that yields constant rewards and a risky arm), it may provide fewer choices than more general MABs do. Although UCB has theoretic guarantees on regrets, it is a deterministic method based on optimistic assumptions. On the other hand, Thompson sampling is fully Bayesian in that it selects an arm based on random samples taken using the posterior distribution (Eq 2) that is updated based on observations. A well-known empirical work [20] also shows Thompson sampling is more robust than UCB is. A discussion of more detailed performance evaluation results follows.

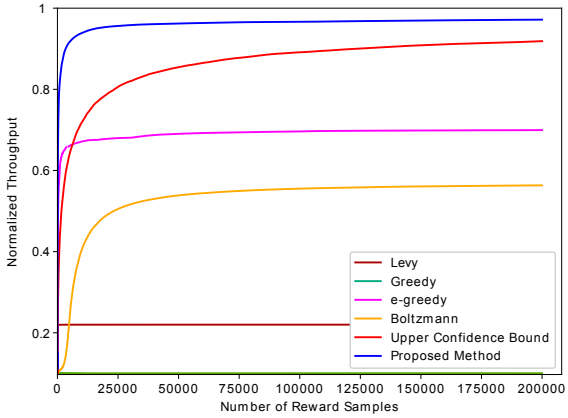

Fig. 3: Comparisons of Optimal Arm Selection Methods

Figure 3 compares the throughput normalized to the maximum possible throughput that can be achieved if complete future knowledge of all path performance is available, which is impossible in practice. As shown in the figure, the achieved normalized throughput of the greedy method [15], the ϵ-greedy method [17], the Levy method [16], the Boltzmann method [18], the UCB technique [1], [2], and our approach based on Thompson sampling in the steady state are approximately 0.01, 0.7, 0.1, 0.5, 0.9 and 0.98, respectively. Thus, our approach enhances the normalized throughput by approximately between 8–97% compared to the tested baselines. It shows that in case of largely varying reward samples, the performance of the baselines degrades significantly. The proposed approach based on MAB via Thompson sampling, however, supports the highest throughput among the tested approaches, because it continuously updates the posterior distribution based on the reward and regret information, while efficiently balancing exploration and exploitation as discussed in Section III.

The UCB algorithm [1], [2], which shows the highest performance among the tested baselines, and our approach are further evaluated for the direct and indirect communication mode in Figures 4 and 5, respectively. In the direct communication mode shown in Figure 4, the normalized

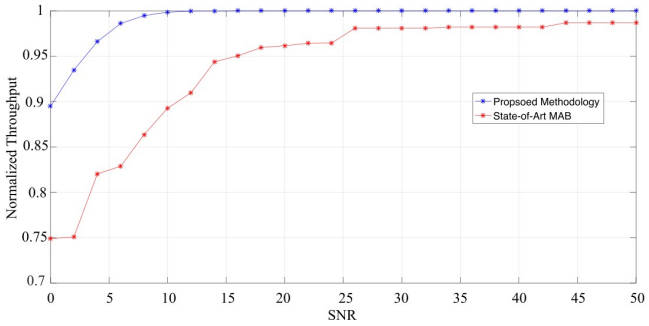

Fig. 4: Evaluation for the Direct Communication Mode

throughput of our approach is higher than that of [1], [2] by up to approximately 15% due to the robustness.

In the indirect communication mode, our methodology outperforms [1], [2] by up to approximately 13% as shown in Figure 5. In both Figures 4 and 5, our approach outperforms the best-performing baseline (i.e., UCB) [1], [2] especially when the SNR is relatively low. This is because the capability of choosing an optimal arm via Thompson sampling in our approach makes bigger differences compared to the UCB method used in [1], [2] when the SNR is lower, since Thompson sampling is not based on optimistic assumptions but is fully Bayesian as discussed before.

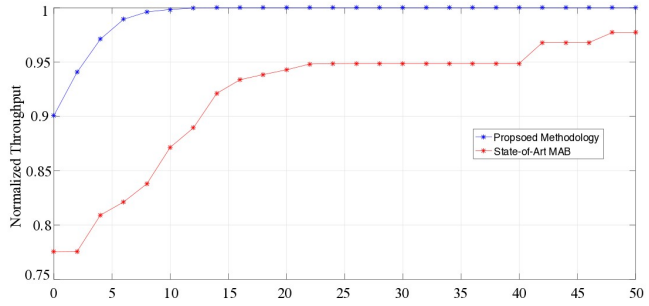

Fig. 5: Evaluation for Indirect Communication Mode

V. Conclusions

In heterogeneous 5G networks, it is essential yet challenging for a user equipment to choose an optimal communication mode and a neighbor to which it connects, e.g., a macro/small base station or a nearby relay/device in the direct or indirect communication mode, respectively. In this paper, we apply the multi-armed bandit technique based on Thompson sampling to pick an optimal neighbor, while effectively dealing with the exploration-exploitation dilemma. In a simulation study undertaken in Matlab, our approach enhances the (normalized) throughput by approximately 8–97% compared to several baselines that represent the current state of the art. Moreover, it improves the throughput by up to 15% compared to the best performing baseline [1], [2]. Overall, research on communication mode selection in heterogeneous networks is still in an early stage; therefore, there is ample room for improvement. In the future, we will continue to investigate more advanced approaches to further enhance the performance in heterogeneous networks.

Acknowledgement

This work was supported, in part, by NSF Project CNS-1526932. We appreciate anonymous reviewers for their help to enhance the paper.

References

[1] S. Maghsudi and E. Hossain, "Multi-armed bandits with application to 5G small cells," *IEEE Wireless Communications*, vol. 23(3), pp. 64–73, 2016.

[2] S. Maghsudi and D. Niyato, "On Transmission Mode Selection in D2D-Enhanced Small Cell Networks," in *IEEE Wireless Communications Letters*, vol. 6, no. 5, 10 2017, pp. 618–621.

[3] Kenan Xu, H. Hassanein, G. Takahara, and Quanhong Wang, "Relay Node Deployment Strategies in Heterogeneous Wireless Sensor Networks," *IEEE Transactions on Mobile Computing*, vol. 9(2), pp. 145–159, 2010.

[4] D. R. Dandekar and P. Deshmukh, "Relay Node Placement for Multi-Path Connectivity in Heterogeneous Wireless Sensor Networks," in *International Conference on Computer, Communication, Control and Information Technology*, vol. 4. Elsevier, 2012, pp. 732–736.

[5] M. Li, L. Bai, Q. Yu, and J. Choi, "Beamforming for Dual-Hop MIMO AF Relay Networks with Channel Estimation Error and Feedback Delay," in *IEEE Access*, vol. 5, 2017, pp. 21 840–21 851.

[6] D. S. Gurjar, S. Member, P. K. Upadhyay, S. Member, D. Benevides, S. Member, and R. Tim, "Beamforming in Traffic-Aware Two-Way Relay Systems With Channel Estimation Error and Feedback Delay," *IEEE Transactions on Vehicular Technology*, vol. 66(10), pp. 8807–8820, 2017.

[7] S. Sun, K. Adachi, P. H. Tan, Y. Zhou, J. Joung, and C. K. Ho, "Heterogeneous network: An evolutionary path to 5G," in *IEEE Asia-Pacific Conference on Communications*, 2015, pp. 174–178.

[8] M. Ali, S. Mumtaz, S. Qaisar, and M. Naeem, "Smart heterogeneous networks: a 5G paradigm," in *Telecommunication Systems*, vol. 66, no. 2. Springer US, 2017, pp. 311–330.

[9] RCR Wireless News, "North American enterprises to deploy 400,000 small cells this year," 2018.

[10] R. Kumbhkar, N. Mandayam, and I. Seskar, "HetNetwork Coding: Scaling Throughput in Heterogeneous Networks using Multiple Radio Interfaces," in *Networking and Internet Architecture*, 2014.

[11] J. Lee and T. Q. S. Quek, "Heterogeneous network throughput with hybrid-duplex systems," in *IEEE Global Communications Conference*, 2014, pp. 3635–3640.

[12] J. White, *Bandit Algorithms for Website Optimization*. O'Reilly Media, 2012.

[13] S. Agrawal and N. Goyal, "Analysis of Thompson Sampling for the Multi-armed Bandit Problem," in *Annual Conference on Learning Theory*, no. PMLR 23 (39), 6 2012, pp. 1–39.

[14] N. Gupta, O.-C. Granmo, and A. Agrawala, "Thompson Sampling for Dynamic Multi-armed Bandits," in *International Conference on Machine Learning and Applications and Workshops*, 2011.

[15] Peishuo Li, T. Vermeulen, H. Liy, and S. Pollin, "An adaptive channel selection scheme for reliable TSCH-based communication," in *International Symposium on Wireless Communication Systems*, 2015.

[16] S. Maghsudi and S. Stanczak, "Transmission mode selection for network-assisted device to device communication: A Levy-bandit approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[17] T. Kato, N. A. Farahin Kamarul Zaman, and M. Hasegawa, "Application of multi-armed bandit algorithms for channel sensing in cognitive radio," in *IEEE Asia Pacific Conference on Circuits and Systems*, 2012.

[18] H. Li, "Learning the Spectrum via Collaborative Filtering in Cognitive Radio Networks," in *Symposium on New Frontiers in Dynamic Spectrum*, 2010.

[19] J. R. Shahid Mumtaz, Ed., *Smart Device to Smart Device Communication*. Springer, 2014.

[20] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," in *International Conference on Neural Information Processing Systems*, 2011.

[21] D. Russo, B. V. Roy, A. Kazerouni, and I. Osband, "A Tutorial on Thompson Sampling," *CoRR*, vol. abs/1707.02038, 2017.