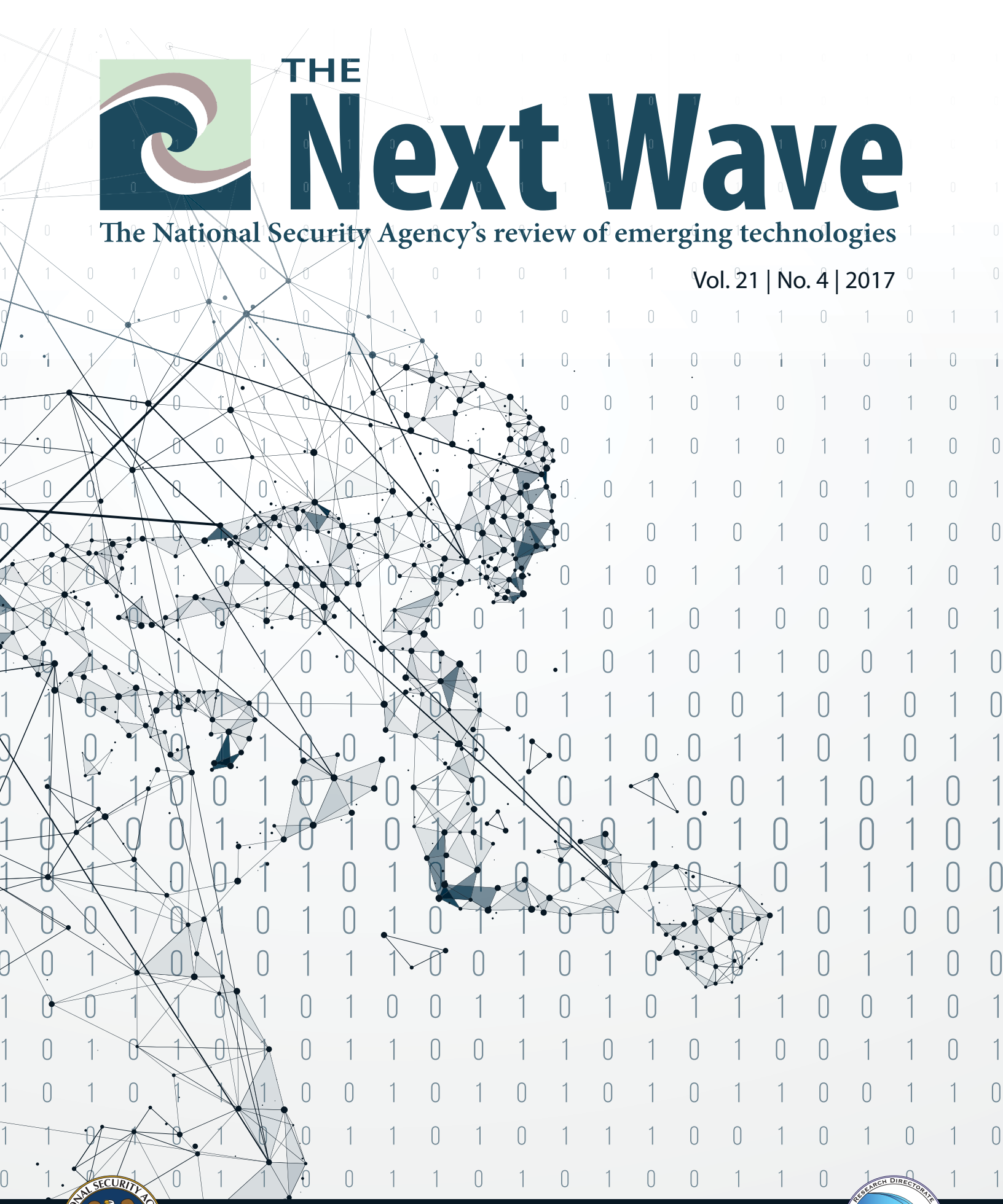
THE Next Wave

The National Security Agency's review of emerging technologies

Vol. 21 | No. 4 | 2017



5G Security & Privacy

[Photo credit: liuzishan, Olga_Rom/iStock/Thinkstock]

THE Next Wave
The National Security Agency's review of emerging technologies

GUEST Editor's column

David James Marcos

The age of ubiquitous computing stands at our doorstep. Barely have we seen the full advent of cloud computing; barely have we fully comprehended the implications of our now mobile-connected livelihoods—yet we are now presented with a world that only recently seemed in the realm of science fiction.

The Internet of Things—coupled with the cloud and enabled with the promise of high-speed fifth-generation (5G) communications—promises to intensify the technological revolution we are already amidst: a place where everything is interconnected, transmitting multifarious data to multifarious systems. What was the cloud is becoming a fog and, some foresee, even a mist. The technology beckoning at our doorstep broaches the cyber-physical, becoming seemingly inseparable from ourselves and the objects around us at a molecular level.

This edition of *The Next Wave* discusses some of the technical obstacles underlying this profound step forward in our increasingly connected lives. Focusing predominantly on 5G technology, the implications of privacy and security are contrasted across the competing requirements of high-speed, low-latency connectivity of billions of devices.

While the articles discussing 5G end-to-end security and LTE Direct predominantly discuss security, the privacy implications are obvious. As low-security devices interact, at scale and speed, with one another directly, indirectly, and with the cloud simultaneously—through both private and public networks—privacy challenges equal if not surpass those of security. The article on models to quantify privacy risk reinforces this point. As is opined in the article, the emerging models, being initial thoughts intended to spur research and discussion, illustrate the predicament of privacy, namely that privacy solutions lag behind security. Contrasting this against the initial points of the 5G end-to-end security article, the privacy predicament is borne out: The article conveys challenges for security-focused monitoring of devices while the potentiality of monitoring for privacy-related issues has not been fully considered. Critically, much research is still needed to refine usable, scalable, and technologically oriented privacy risk quantification.

The privacy and security considerations in the age of ubiquitous computing are further compounded by the ever-increasing regulatory requirements imposed upon information technology. The European Union's General Data Protection Regulation (GDPR) stands as the most salient and current example of such. Raising *data protection* to the status of a human right, the GDPR levies significant financial penalties on companies that do not provide adequate protection, notice, consent, and redress for consumers. This paradox—the increasing connectivity and ubiquity of our

personal data and devices contrasted against increasing regulatory strictures—proves out the continuing need to develop rigorous, scalable, scientifically verifiable security and privacy solutions; solutions that at once safeguard technology while minimizing the real harms to our privacy, civil liberties, and livelihoods posed by a seamlessly connected future.

NSA's Research Directorate is and remains fully committed to such, notably through the ongoing efforts of the Science of Security and the nascent efforts of the adjacent and overlapping Science of Privacy. In a world where computing is ubiquitous, where a mist of data and devices diffuses into our lives, where that mist becomes inseparable—indistinguishable—from reality, trustworthy computing is but axiomatic.

David James Marcos

Privacy Research Lead,
Information Assurance Research, NSA

Contents

The Next Wave is published to disseminate technical advancements and research activities in telecommunications and information technologies. Mentions of company names or commercial products do not imply endorsement by the US Government. The views and opinions expressed herein are those of the authors and do not necessarily reflect those of the NSA/CSS.

This publication is available online at http://www.nsa.gov/thenextwave. For more information, please contact us at TNW@tycho.ncsc.mil.

Vol. 21 | No. 4 | 2017

Investigating End-to-End Security in 5G Capabilities and IoT Extensions*

Jason J. Uher (Johns Hopkins Applied Physics Laboratory; JHU APL), Jason R. Harper (JHU APL), R. G. Mennecke III (JHU APL), Pamela M. Patton (JHU APL), Dr. Sam Farroha (Laboratory for Telecommunication Sciences)

The emerging fifth generation (5G) wireless network will be architected and specified to meet the vision of allowing the billions of devices and millions of human users to share spectrum to communicate and deliver services. The expansion of wireless networks from its current role to serve these diverse communities of interest introduces new paradigms that require multitiered approaches. The introduction of inherently low-security components, like Internet of Things (IoT) devices, necessitates that critical data be better secured to protect the networks and users. Moreover high-speed communications that are meant to enable the autonomous vehicles require ultra-reliable and low-latency paths. This research explores security within the proposed new architectures and the cross interconnection of the highly protected assets with low-cost, low-security components forming the overarching 5G wireless infrastructure.

1. Introduction

According to the International Telecommunication Union (ITU) and its vision for the fifth generation (5G) of cellular [1], there are going to be three main drivers for 5G: enhanced mobile broadband, low-latency ubiquitous connectivity, and machine type communications.

The first driver is of no surprise; each generation of cellular since 2G has brought with it an increase in data speeds to the user, and each time, the consumer demand for that data has exceeded expectations. With the ever-increasing user appetite for data, 5G will be no exception—the ITU is mandating a rate of 20 gigabits per second (Gbps) peak data rates and 100 megabits per second (Mbps) minimum to the user for their 5G specifications, rivaling even most wired home networks. From a security perspective, this raises a number of interesting, though not insurmountable, challenges. The use case for enhanced mobile broadband (eMBB) is largely understood, and even though monitoring such large data connections has its challenges, it's a matter of scaling current technology to meet that demand. This will not be a simple problem to solve, however. With the current level of security technology, 5G data throughput would require massive server farms running intrusion scanners and packet inspectors just to keep up with the nominal state of a network. In addition, signaling for connection setup between heterogeneous connections and handover in distributed data networks will make session monitoring difficult. While it is clear that eMBB

*Article republished with permission from SPIE: Proc. SPIE. 9826, Cyber Sensing 2016, 98260A. (May 12, 2016) doi: 10.1117/12.2229608.

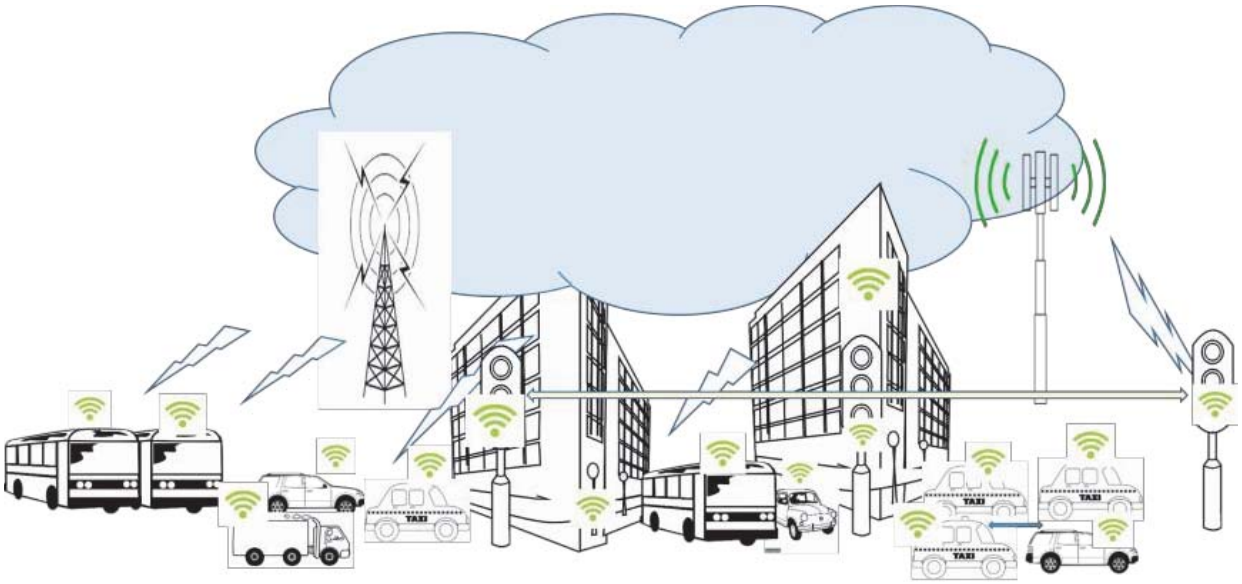


FIGURE 1. Illustration of device-to-device connectivity for vehicular applications.

is the simplest security case for 5G, even it will not be a simple undertaking—much work will be required to ensure that users and networks can be kept safe despite the huge amounts of data that need to be carried at breakneck speeds.

In addition to increased data speeds, the ITU has highlighted machine type communications (MTC), or "Internet of Things" (IoT), and low-latency ubiquitous computing as target use cases for 5G networks. Introducing these new use cases signals a move away from the traditional mobile telephony model and complicates the network security significantly. Ensuring a secure network for these new models will require not only new technology, as with eMBB, but entirely new paradigms for how security is considered within the network. IoT networks, for example, will require large numbers of devices to communicate with each other, likely without a central coordinator. In addition, IoT use cases include things like self-driving cars, which will require methods for communicating vehicle-to-vehicle (V2V) and even vehicle-to-infrastructure (V2X). Figure 1 illustrates the concepts of connecting vehicles to achieve the autonomous vehicle vision. These new communication models will introduce a whole host of new paradigms that will not fit within the traditional security models of mobile telephony.

For the second new use case, low-latency communications, ensuring a low-latency link will severely limit the amount of observation time the network has and reduce the length of time packets can be analyzed. In short, current network technology will be so busy ensuring the latency requirements for a 5G network that there will be no time or processing resources left over to scan the traffic for security threats. While the mandate for a packet switched core in 4G has started the move towards the data-centric security models, the MTC and low-latency use cases will introduce entirely new security models to the mobile telephony space. Driverless cars that can communicate with each other promise a safety revolution by warning each other of danger ahead and start breaking with superhuman reactions. However, current networks are unreliable and can only produce a minimum of ~40 milliseconds (ms) of latency.

1.1 What is "end-to-end" security?

Current network security models, both in enterprise deployments and the mobile telephony realm, rely on a minimum amount of responsibility distribution, with a central point of control and authority to do monitoring and tasking. Current models largely

Acronym List

LOS	line-of-sight
MAC	media access control
Mbps	megabits per second
MiTM	man-in-the-middle
MME	mobility management entity
MMIMO	massive multiple input, multiple output
mmWave	millimeter wave
ms	millisecond
MTC	machine type communications
NFV	network function virtualization
NOC	network operations center
OFDM	orthogonal frequency division multiplexing
ONOS	open network operating system
PDP	packet data protocol
PHY	physical
QAM	quadrature amplitude modulation
QoE	quality of experience
QoS	quality of service
RAN	radio access network
RAT	radio access technology
RNC	radio network controller
SDN	software-defined networking
SDR	software-defined radio
SDWN	software-defined wireless networking
SeGW	security gateway
SGSN	serving general packet radio service support node
SIM	subscriber identification module
SNR	signal-to-noise ratio
TMSI	temporary international mobile subscriber identity
TOSCA	topology and orchestration specification for cloud applications
UE	user equipment
URC/LL	ultra-reliable low-latency
uSIM	universal subscriber identity module
V2V	vehicle-to-vehicle
V2X	vehicle-to-infrastructure
VM	virtual machine
VNF	virtualized network function
WSN	wireless sensor network

consist of the "observer" and the "controller" roles in a given system. Observers monitor something within the network and report it back to the central controller, which makes security decisions with a global view. For example, a router may be observing traffic and reporting connection information back to a Network Operations Center (NOC), which aggregates that with information from all the other routers and displays any suspicious activity to the operators. This model will simply be untenable in 5G networks. Every single thing that might need to be monitored by a central entity will scale to unmanageable numbers with the new services—eMBB will increase the total amount of data, IoT will increase the number of connections, and the low-latency communications will limit the amount of traffic that can be inspected. In the best case scenario for a 5G central controller, data can be scanned after the fact for threats, and even then, offline processing speeds will require a dramatic upgrade from today's technologies to ensure the average output rate doesn't exceed the input rate.

While it is clear that the current centralized model of network security is not going to work for 5G, it remains unclear what can be done about it. It is obvious that a global view of all traffic will lead to optimal decisions regarding the security of the network, but the sheer amount of data and connections in a 5G network will make that impossible. It is also clear that security functions will have to be distributed throughout the network and, as often as possible, handled locally by the processing elements in the network, not a central controller. This distribution of security responsibility is going to introduce a number of issues that stem from both the increased data rates of eMBB and from the adoption of new use cases within 5G.

Each of the following sections in this paper highlights a number of areas that will require scrutiny in the new 5G security models. The next two sections of this paper, Sections 2 and 3, focus on the fundamental source of the security concerns in 5G: system aspects, new technologies, and new paradigms. The first section focuses on the security impact to the entire 5G ecosystem that results from the blending of multiple, disparate use cases. The next section covers how new technologies proposed for 5G will impact existing security models in ways that may introduce new security concerns or alleviate old ones within the traditional telephony systems. Finally, the sections consider how

new paradigms will drive a completely new set of security issues not previously considered in the mobile telephony space.

Though separating underlying technologies from the network use paradigms is useful for analysis, the reality is that they will remain forever intertwined during implementation. This is why potential solutions to these new issues have been broken out into a fourth section which discusses the application of new security methods to the problems presented in the three preceding sections.

Finally, a word of caution: This paper is intended merely to highlight the open areas of security research within the looming 5G standards definitions. No single paper can encompass all of the security implications of a new, as yet undeveloped, standard. This paper focuses on the issues that we feel are the biggest threat to the security and privacy of users within future 5G networks and proposes high-level solutions that may be considered for each of the identified issues. We do not intend that this paper will serve as a comprehensive source of 5G security implementations, nor that our presented solutions will provide the greatest overall safety for the network. There is much work yet to be done in 5G security.

1.2 System-level considerations

Of all the security issues that will eventually be raised when developing the next 5G standard, the most vexing will almost certainly be how to reconcile, at a system level, the conflicted nature of the ITU's International Mobile Telecommunications (IMT) vision. The three primary use cases, eMBB, machine type communications (MTC/V2V/V2X), and ultra-reliable, low-latency communications (URC/LL), all have specific requirements from an implementation perspective that places them at odds with each other. Figure 2 shows a graphic from the IMT 2020 Vision

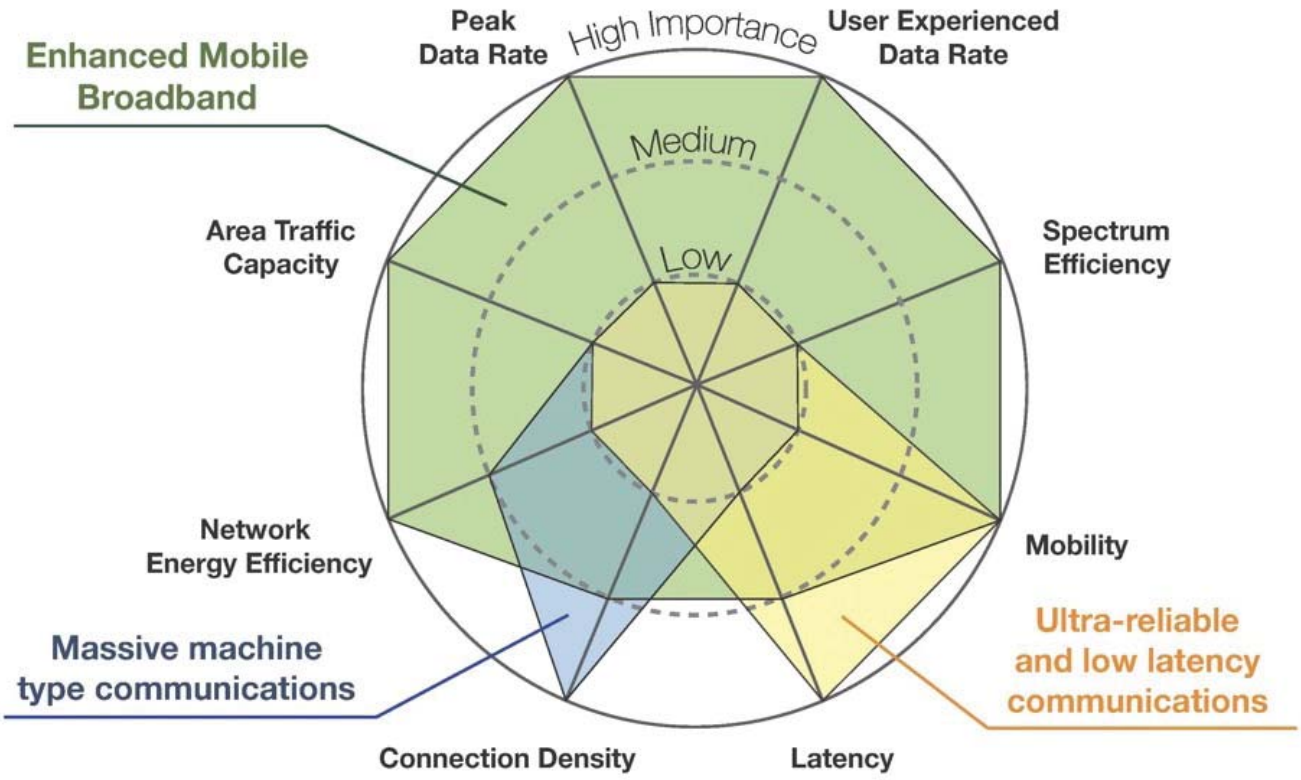


FIGURE 2. Eight key capabilities and their relative importance to 5G use cases [1].

paper that directly demonstrates this conflict—parameters that are important for each of the goals are almost exclusively not important for the other two.

When considering the requirements for the three main use cases that IMT has identified for 5G, the most important are those related to MTC and URC. The eMBB use case is certainly going to be the driver in terms of initial standards development; it is likely the fastest way for the carriers and equipment manufacturers to recoup their investment (through access fees) and will likely be the largest segment of their customer base on a per-account basis. What is interesting, though, is that the other use cases are far more likely to have an impact on the network as a whole. On a *per-device* basis, MTC will likely far outstrip the eMBB customers in number of connections, URC will require large amounts of bandwidth during poor channel conditions, and low-latency communications will utilize a large percentage of any high-speed switching fabric within the core. This implies that even though the carriers will likely see eMBB as the most important feature, the MTC and URC/LL use cases will be what ultimately determine whether their network meets the requirements for 5G.

From a security standpoint, some of the apparent conflicts may be resolved easily. While there are specific issues related to low-latency communications, as presented in section 4, low latency, URC, and eMBB connection patterns will be nearly identical. This leaves MTC as the barrier to a full "system level" scheme for handling security. It is likely that symmetric and asymmetric encryption schemes will be proposed in 5G for both data integrity and privacy. In 4G and 5G cellular services, symmetric key derivation and generation lies at the heart of the security paradigm: Everything rests on a single secret shared between the network operators and the universal Subscriber Identity Module (uSIM) in the phone. All connections are built upon this shared secret, and all authentication between the user and the various parts of the network rely on generating keys from this shared secret. In 4G LTE, this key generation and exchange process is nontrivial; several messages must be sent between the user and a security agent at each level of the networking stack. With the connection densities proposed for a 5G MTC system, the packet core would be quickly overwhelmed, doing nothing but trying to authenticate devices as they come on to the network and are periodically re-keyed. Clearly, MTC requires a new way of thinking about how user authentication and attachment is handled in the network.

The issue of consolidating disparate uses cases into a single waveform is currently under discussion in both the Institute of Electrical and Electronics Engineers (IEEE) [2] and Third-Generation Partnership Project (3GPP) [3] from an implementation perspective. Even without considering security, it is unclear how the conflicting constraints can be met within a single waveform. At present, there seem to be two competing strategies—radio access network (RAN) slicing and flexible waveform numerologies.

The first strategy, called RAN Slicing, is to simply ignore the problem altogether and use separate radio access technologies (RATs), or 'slices,' for each of the three use cases. This allows the radio signaling to be as fast or slow as necessary, and allows traffic within the packet core to be largely treated as use case-agnostic data [with quality of service (QoS) to ensure the low-latency requirements when necessary]. The primary issue with separate RATs is that resources may not be dynamically allocated to the different use cases as demand changes. When specific frequency resources are devoted to a particular RAT, it is not likely that they can be easily shifted to another. For example, if the number of broadband users in a given area decreases, it is unclear that the spectrum resources can be reallocated in order to decrease the reporting intervals of a sensor network.

In the second case, flexible waveform numerologies, the network would use a system similar to the current 4G LTE standard. At regular intervals, frames containing a combination of control data, user downlink data, and user uplink data will be scheduled, allowing the network and each user to transmit in their own time/frequency allocation. Through clever assignment of orthogonal frequency-division multiplexing (OFDM) numerologies, frames can be allocated such that an MTC device may use a low bandwidth, unauthenticated resource block in one resource block while the eMBB customer receives their downlink in the next, and the URC/LL subscriber is constantly getting blocks in each frame.

From a security perspective, the RAN Slicing approach is compelling—it allows for the physical layer encryption and authentication to be handled

completely differently for the eMBB, URC/LL, and MTC waveforms. While a future standard could, theoretically, allow for different authentication and data protection mechanisms utilizing the flexible numerology architecture, it will be difficult to ensure that control data is appropriately protected unless the same protection mechanism is used across all frames. This means that any protection mechanisms must be sufficiently simple that battery-constrained IoT devices can perform the cryptographic operations, or that the low-latency devices can very quickly decrypt and verify the data. This leads to a scenario in which substandard protection mechanisms may be used to protect the privacy and confidentiality of control data in favor of a unified RAT.

In the end, the system level considerations for a future 5G system boil down to whether or not a unified schema can be found to unite the conflicting technological and social requirements of the different use cases. A multi-RAT solution may be able to provide each use case with independent methods for security, alleviating the need for such a unified architecture, but it is more likely that a single RAT will be developed to give networks flexibility in planning their operations. If this is the case, security must be considered from the beginning; building a security model based on assumptions driven only by the eMBB use case will almost certainly result in subpar security performance in the IoT and low-latency use cases.

2. New paradigms for 5G networks

With each new generation of mobile telephony standards, security has become an increasing concern. The most recent 4G standards now ensure that both the network and the users have guarantees on mutual authentication, privacy, and message integrity throughout the network. However, up until now, the models for network architecture have remained largely the same: a central authentication authority works with a distributed network of radio heads to provide voice and data service to user equipment. As a result, the security models in use today are highly tailored to this specific use case. Both the network architectures and security models rely on the assumption that user data is going to be asymmetric in bandwidth on the uplink and downlink, and that the data endpoints will generally lie outside of the core network, such as on the Internet.

New use cases presented for 5G adoption will greatly challenge this traditional security model. New paradigms, such as disconnected operation, small cell data links, edge-focused processing, and more, will turn the central authority authentication model on its head. In this section we call out some of the big problems facing 5G security models, citing examples of technologies that are not supported by the central authority model. The first part of this section covers new security implications in small cell deployments and the issues that arise when user data is spread across multiple concurrent connections. The next part of this section provides an overview of the security considerations when devices are allowed to connect directly with one another, either with assistance from the central network or while in a disconnected state. Finally, we consider the MTC, or IoT, paradigm. Due to the extremely dense deployments and relatively limited power budgets, this new use case has the highest probability of disrupting the current authentication models of mobile telephony services.

2.1 Concurrent multilink approach

The multilink capability will allow multiple simultaneous links between mobile devices and the rest of the network including cellular towers, Wi-Fi base stations, and other mobile devices. The approach utilizes licensed and unlicensed spectrum to improve capacity of the mobile networks. The unlicensed spectrum, most likely in the 5 gigahertz (GHz) band, can be leveraged either using existing technologies in that band (Wi-Fi) or by modifying some cellular technologies to aggregate those channels into the emerged cellular physical (PHY) and media access control (MAC) layer. The approach is causing disagreements among cellular and Wi-Fi ecosystems regarding its impact on existing and future Wi-Fi networks. There are many prototyping efforts via realistic test beds to understand the performance trade-offs. The evaluation includes comparing performance trade-offs such as interference, algorithm implementations in realistic environments, security, and robustness. The results should drive the architecture of future networks.

2.2 Small cell technologies

As cellular networks have continued to evolve from the voice use cases to high-data throughput

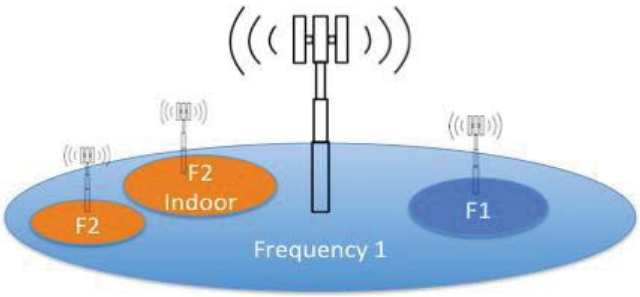


FIGURE 3. Small cell scenarios.

applications, cellular standards have evolved to address this need. Not only have the cellular standards evolved for these use cases, but also the need for different cellular topologies are being introduced based on where users consume this high-data throughput, such as event venues, offices, or customer premises. To address this paradigm shift, the introduction of small cell technologies and concurrent multilink networks has emerged as a viable solution. Small cells are being used and considered for increasing both coverage density in a region and for addressing poor reception for either cell boundary conditions or places where the signal-to-noise ratio (SNR) is severely degraded, such as indoors. By bolstering coverage in a particular area, there are now more resources to serve the users. In addition, by bringing the small cell indoors or closer to the end-user device, one can expect to achieve a higher SNR. Within LTE Release 12, small cells will begin to take advantage of this higher SNR by offering 256 quadrature amplitude modulation (QAM).

Small cells are expected to work in a variety of different scenarios illustrated in figure 3. These scenarios include cases with and without the macro cell present and where the small cell is located on the same or different carrier frequency from the serving macro cell. Within these network topologies the user equipment (UE) can be served by an individual cell or by a combination of cells. Within the context of multilink connectivity, the UE can be served either by both the macro cell and the small cell using Coordinated Multi-Point (CoMP), first introduced in LTE Release 11, or through dual connectivity, first introduced in LTE Release 12. The main difference between these two configurations is that in the latter, the UE operates two MAC entities. These small cells can be deployed in the traditional form where the carrier installs the equipment at a location they own and operate or at a customer's home [4].

2.2.1 Small cell security

Small cells have been and will continue to be deployed into cellular network topologies, and while there are great benefits to both the carriers and the users, these come at the cost of additional complexity and security concerns.

2.2.1.1 Small cell authentication

As entry points into the carrier's core network, it is of utmost importance that the network properly authenticates the small cell. In LTE networks the small cell connects back to the Security Gateway (SeGW), which has the task of authenticating the small cell onto the network. After authentication, the small cell is provided its frequency allocation and other network parameters from the core network. Given that this process is done either at a customer's home through their network or a carrier's private network, encryption of this process is mandatory.

2.2.1.2 Device tampering

While the traditional macro cell is contained at a secure site owned by the carrier, that is not the case in all small cell deployments. These small cells can be located within office buildings, event venues, or even a customer's home. Due to the device being easily accessible, device tampering must be considered. Open ports and debugging interfaces all pose as potential access points to gain unauthorized access to the device.

2.2.1.3 Network trust

Small cells placed at the customer's home cause another security concern. This device would be connected up to the customer's home network, which provides an interface that the customer does not fully own or have the ability to verify. The device is provided by the carrier and should be secure, but the customer has little to no insight into virus protection, security vulnerabilities, and the application of software patches. If not properly secured, this device provides access to the customer's home network behind their firewalls.

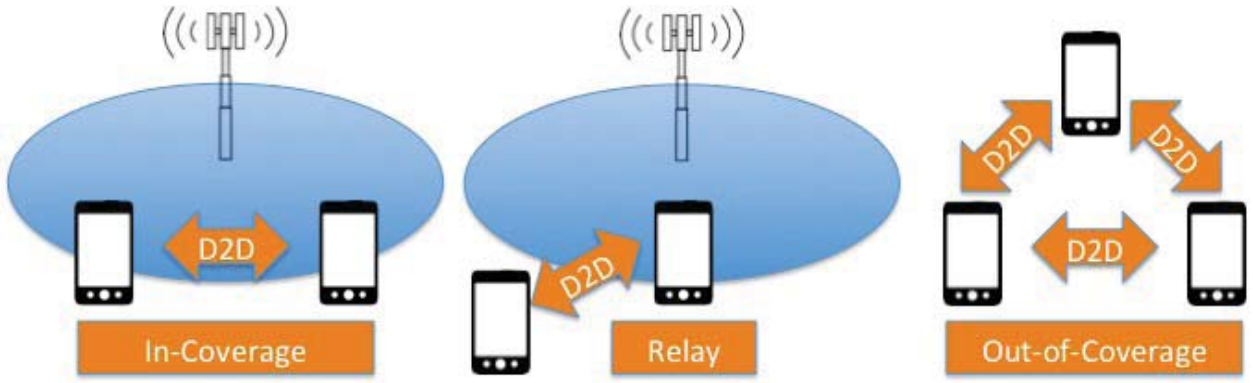


FIGURE 4. LTE Release 12 D2D use cases.

The network provider also assumes a great risk in allowing these small cells to connect to the core network through the Internet. By providing this access through the SeGW, the small cells and the SeGW are susceptible to denial-of-service (DoS) attacks. This could limit data rates and even access of one or more small cells to provide the expected QoS.

2.2.2 UE security

While there has been more emphasis placed on UE security, the UE is often overlooked. This has improved with LTE's mutual authentication and should continue into 5G networks.

2.2.2.1 UE tracking

Much interest has been shown in International Mobile Subscriber Identity (IMSI) and Temporary Mobile Subscriber Identity (TMSI) sniffing and tracking from the Global System for Mobile Communications (GSM) networks onward. With the addition of small cells, this will continue to be a problem but ultimately could increase the potential effects of these trackers. By observing the signaling of a small cell or cells, one could track a particular node to a small cell, and with the cell's limited footprint, the tracker would have greater location accuracy. If this capability was to be spread across multiple small cells, this could be used for movement tracking of a particular user.

2.2.2.2 Small cell trust

Given the security concerns discussed, the UE should be concerned with either rogue small cells or compromised small cells. A UE, if connected to a malicious small cell, could be redirected or be served intentionally corrupted data.

2.2.2.3 Dual connectivity security concerns

In the CoMP case, a UE places its trust in the serving macro cell and allows the network to coordinate and provide other resource allocations from neighboring cells. The UE does not authenticate the neighboring cell. In the dual connectivity case outlined in LTE Release 12, the UE has two MAC instances running and authenticates with both the main cell and the secondary cell. A split data pipe from two cells requires that the UE has the additional burden of maintaining two links with additional overhead signaling and verifying their integrity as well as the data that they are serving.

2.3 Device-to-device communications

As more and more devices are connected to the network and social applications and additional use cases emerge where nearby devices communicate, migration of that data from traversing the core network to a means where these devices communicate directly with one another is ideal. The 5G standards bodies are exploring how these device-to-device (D2D) communications will materialize. Use cases involve data offload, range extension, and proximity services including advertising, V2V, and vehicle to everything. Within LTE Release 12, D2D has been defined for three use cases (shown in figure 4)—in-coverage, relay, and out-of-coverage—but will continue to expand in 5G networks [5].

2.3.1 What does it mean when devices are talking directly to each other?

Devices directly talking to each other will be established with or without network interaction. The in-coverage scenario allows for the devices to be able to discover one another and establish their D2D link fully coordinated by the serving cell. In other cases, the UEs will require additional features to assist in discovery and D2D communications setup. With devices now directly communicating with one another, there are security concerns in the authentication of the UE that will be in direct communications. This direct communications link could be vulnerable to impersonation or playback attacks. This link provides a direct path for a UE to be interrogated by another UE on the network. This interrogation could compromise their identity, location, or other information about the user.

In the case of relay nodes, both the relay node itself and the end UE are vulnerable. The far-end UE must be able to identify and verify the relay UE node and create a secure tunnel to the serving cell through the UE as its data path. This link is susceptible to man-in-the-middle (MiTM) attacks, where the far-end UE and the network need to be able to identify compromised relay nodes and be able to verify the sender and integrity of the data being exchanged. The far-end UE could pose a threat to the relay node by heavily utilizing its resources and draining its battery life. The relay in this scenario must not be addressable by the far-end UE.

2.3.2 Disconnected edge communications

Without network involvement, establishing these D2D links becomes a security challenge. Within the context of LTE Release 12, the only D2D use case without network interaction is for public safety. In this case, users would be provided devices that have additional capabilities in the devices themselves or activated through special Subscriber Identification Module (SIM) cards. As we progress towards the 5G use cases to include V2V, it would not make sense to deny a V2V exchange that could prevent an accident. There are security concerns with how to authenticate another device, whether to have keys that the network providers place in the SIM card, rolling keys provided by the network during the device's last connection, caching previous D2D connections, or through distribution with other trusted devices. Each of these cases provides vulnerabilities to be exploited. In cases where the network is not present, the network provider could request logging of D2D interactions to later verify the users and blacklist malicious devices from future D2D interactions.

2.4 MTC and the IoT

2.4.1 The cloud, the fog, and an IoT

Cloud computing has moved the computing and storage resources from the office to the World Wide Web. With this transition, users are now relieved of the responsibilities of information technology infrastructure management and can focus solely on the computing aspect of their business. Cloud computing also shifts the security posturing to a more centralized location so that more resources can be used to secure the cloud and access to it. As a financial incentive, cloud computing offers a pay-as-you-go resource utilization paradigm that ensures only necessary resources are used, which is a cost-saving optimization for resource allocation and infrastructure management.

2.4.1.1 The Cloud of Things

At first glance, cloud computing looks like a reasonable solution to meet the demands of the IoT. There are different forecasts for the growth of IoT devices on the market; most estimate that there will be 10 to 24 billion devices by 2020. With such a drastic increase in connected devices, there will come an immediate need for storage and processing resources, which cloud computing offers. Localized storage and processing will no longer be feasible in several IoT scenarios, which will push this demand to cloud computing. In this scenario, IoT devices from various heterogeneous networks and domains would be connected and integrated to the cloud as a Cloud of Things (CoT) [6], and all data would flow into the cloud for big data analytics and data processing (as seen in figure 5). The data would then be available to the user as a service or cloud application. There are several factors to consider when integrating IoT into the cloud. From the perspective of cloud security and privacy, by introducing a mixture of private and public IoT data streams for analytics, there will be an added need for data management and segregation of these hybrid

clouds of information. This mass aggregation of private and public data will also lead to an increase in cyber threats and attacks in order to gain access to the data. In order to bring the IoT data to the cloud there will be a need for standardized IoT gateways [6, 7] that can handle multiple IoT protocols, service discovery, identity management, QoS, security, and resource allocation for the cloud.

2.4.1.2 IoT fog networking

The fog networking architecture will play a significant role in the 5G cellular infrastructure and IoT because of the low-latency requirement and the reduction of overall network throughput. Low latency and reducing network throughput are some of the main driving forces behind fog networking and IoT integration. The fog architecture extends the cloud to the edge of the networking by providing localized storage, data filtering, data analytics, and end-to-end communication. This is done by creating localized smaller form factor fog data centers at the edge of the network (local network operation centers). These fog networks are heterogeneous by nature because of the different protocols and device gateways that they must interact with. The fog heterogeneous model draws similarities from the standardized IoT gateway routers needed that will provide a smaller scale version of a fog network, referred to as a mist network, for even more localized IoT deployment. A mist network will consist of minimal localized shared storage, data processing, and a standardized IoT gateway to handle the lowest form of latency in a small form factor IoT network, for example, a home IoT security system or Body Area Network (BAN) using a smart mobile device for processing.

2.4.1.3 Cloud and fog integration (CLoG)

In order to incorporate IoT into the next generation of the cellular infrastructure, 5G IoT will need a cohesive communication standard from end to end. The term CLoG is proposed to describe an end-to-end solution that includes both cloud and fog computing to reach full IoT integration into the new evolution of the Internet and 5G cellular infrastructure. A CLoG system would incorporate smaller independent fog networks that are localized at the edge of the network to provide proper QoS and processing of large data

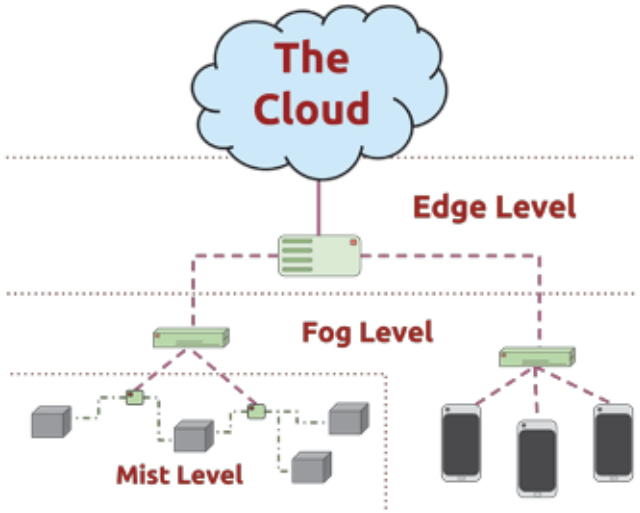


FIGURE 5. Fog network [8].

sets. The fog network can handle resource allocation, service discovery, identity management, and endpoint security. The data can be filtered, and asynchronous metadata can be produced to push to the cloud so that data analytics can provide users with IoT applications as a service when users are in a mobile environment. For more timing-constrained applications, the IoT as a service can be provided at the fog level or mist level, depending on the constraints to the protocol.

2.4.2 Wireless sensor networks

2.4.2.1 Multi-hop mesh topology

In distributed wireless sensor networks (WSNs), there is a need for optimized energy efficiency to reduce power usage while maintaining a robust area of coverage. A typical multi-hop mesh WSN would consist of multiple sensor nodes that might not be directly connected to the hub device or gateway. The nodes can reach the hub by using shortest path multi-hop algorithms to route the sensor data through other nodes to reach the hub or gateway. The advantage to mesh networking is that the sensor network is more robust from node failure because the data can be rerouted or flooded to other nodes to push the data to the central node. Even though power consumption is low in transmission from one node to another, the overall power consumption of the sensor network increases as the amount of sensors increases when the

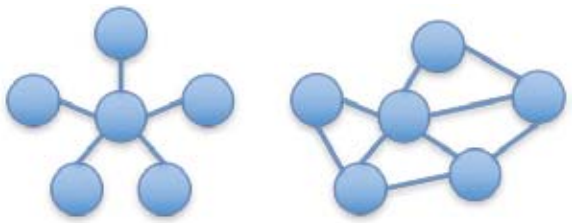


FIGURE 6. Star and mesh network.

sensors at the edge of the coverage area transmit data that must travel through all nodes to reach the hub. In this case, as the sensor coverage increases, so does the power consumption model of the system; this will lead to faster sensor battery replacement scheduling or improvements in the energy harvesting system.

2.4.2.2 Star topology WSN

In contrast to mesh networks, a star network topology consists of multiple sensor nodes that are connected through a central hub. In this topology, each node can reach the central node directly. This cuts down on the multi-hop power consumption that is realized in mesh networks. The main advantage is that the WSN is not affected by a node failure as long as it is not the central node that has failed. The drawbacks to the star topology are seen when sensor nodes are further from the central node or when the sensor network is increased. Longer distances increase transmission power and increased sensor nodes require more connections, which increases power consumption.

2.4.2.3 Ad-hoc star/mesh topology: Bluetooth Low Energy

A third topology is a hybrid that combines attributes from the star and mesh network topologies to form ad-hoc networks, which can be realized in Bluetooth Low Energy (BLE) WSNs. An optimized WSN that combines star and mesh topologies can be realized using a Bluetooth Low Energy (BLE) WSN [9]. BLE devices have a master-slave relationship where each node can be either the master or slave, depending on who initiates the connection. A master node can create a piconet of up to eight sensor nodes (including the master node); a slave node can only be connected to one master node at a time. However, piconets can connect to form scatternets [10] by sharing a slave node that communicates by switching between the piconets and routing communication. This BLE ad-hoc network topology offers the benefits of multi-hop routing for large sensor networks by creating smaller cell piconets with routing capabilities to the main network and across multiple piconets/scatternets. This approach extends the sensor coverage by adding more decentralized, controlled small cell piconets while maintaining the power savings of mesh networks by utilizing the routing methods of the BLE ad-hoc network. The routing will allow all nodes to be equal distance and will thus reduce the power required for transmission and hop processing. Control in this topology is decentralized and distributed across multiple master node piconets. This not only reduces inter-piconet communication by having fewer hops in routing but also makes the WSN less resistant to node failure or control hub failure.

2.5 IoT security concerns

The fog, CLoG, and mist networking paradigm shift has some serious concerns for security. In the smaller form factor, mist network multiple ubiquitous devices will autonomously connect to the IoT gateway. The IoT gateway will be the first frontline of defense against cyberattacks. Some IoT devices will have little or no human interaction, so device identification will be a hard problem to solve. Unless strict device identification standards and certifying agents are implemented across all protocols, device imitation and spoofing will be a common problem to gain access to the IoT gateway. Once the IoT gateway is compromised, the smaller mist network could be leveraged for DoS attacks on the fog node, or the IoT gateway of the mist network could be used to sniff and collect the private data traversing the mist network, while also collecting other sensor device information and identification. In instances where the mist network does the end-to-end authentication with the IoT devices, the data would be decrypted and possibly captured, which would render the IoT encryption method useless. Device spoofing has far-reaching ramifications and implications depending on the IoT system being attacked. For instance, in a home security system, detection sensors could be spoofed to either set off alarms and alert authorities or send the OK signal

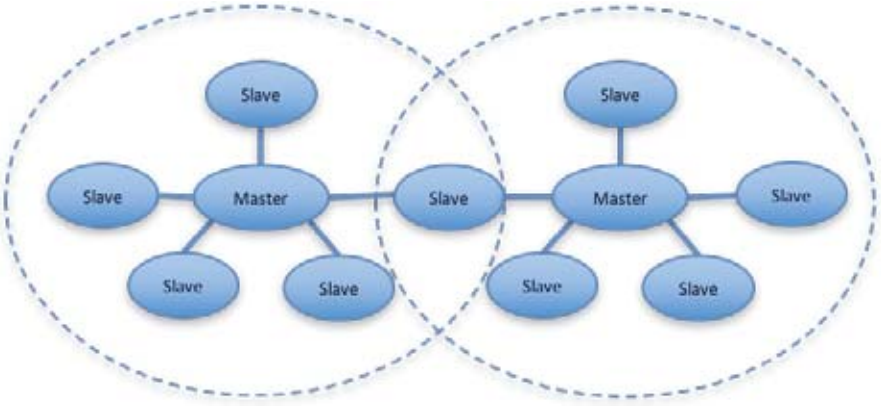


FIGURE 7. BLE ad-hoc network.

while disabling the actual sensor. Device identification could also lead to DoS attacks on sensor networks, and data injection to disrupt IoT services or crash IoT data processing algorithms.

There is no standard encryption method for the various IoT communication protocols in use or proposed. Some IoT protocols have encryption methods but will establish communication at the lowest encryption supported by both of the devices. In that instance, a device with no encryption could gain access to connect to the IoT gateway by simply deescalating its supported security methods. The IoT gateway will need to be hardened to enforce the toughest encryption methods available for each protocol. IoT gateways that are maintained by the user will have more security vulnerabilities because of the unknown security risks of the local network that it might have to connect to. If an IoT gateway must be connected to a local router in order to route data to a fog or cloud network, then that IoT gateway will have a higher risk for cyberattack because of the unknown state of the local router and network. It would be impractical to replace the IoT gateway every time an update is needed, so firmware upgrades are a must. With firmware upgrades on untrusted networks, there is a risk of firmware modification and implantation that can help the attacker gain access to the fog node. The fog node has the same security concerns as the mist network when it comes to device identity, authentication, and encryption.

Within WSNs, the problem of node spoofing can be detrimental. On sensor networks designed with tight power consumption requirements based off of modeling, imposter nodes can be introduced to flood the networks with routing requests or data floods to cause a power-drain attack to render the network inoperable over time. Some methods have been introduced to detect anomalies on networks and blacklist devices suspected of malicious activity. In a sensor spoofing attack without valid device identification, the method used to blacklist the anomaly can be used to blacklist nonthreatening sensors on the network, which could reduce or remove the sensor coverage area.

3. New technologies for 5G

Even without taking new paradigms into account, there will be a number of challenges and effects in 5G security that arise from the adoption of these new technologies. These challenges arise by the nature of subtle changes in the underlying technologies that present a new set of security concerns when applied to old telephony models. For example, the move to use massive multiple input, multiple output (MMIMO) and millimeter wave (mmWave) technologies will potentially increase user privacy, but might simultaneously allow for rapid identification and geolocation of terminals. Similarly, moving to the use of network function virtualization (NFV) inside the core network will reduce operator expenses and streamline operations, but may introduce a number of security flaws that stem from reliance on a centralized core management architecture. In this section, we break down the candidate technologies for 5G with an eye towards required changes in the current security models from the traditional mobile telephony space.

3.1 5G core network architecture

As the RATs evolve to 5G, so must the core network architecture evolve to accommodate increasing bandwidth demands, reduced latencies, and stringent QoS/quality of experience (QoE) requirements. The new core must be flexible to incorporate heterogeneous technologies as well as scalable to quickly add new

capabilities and capacity as needed. Current deployments of 3G/4G networks are based on network appliances that are hardware-centric, with vertical service integration, protocol-specific implementations, and hierarchical connectivity. 5G networks require an evolution to virtualized cloud-based network components providing universally accessible services, application programming interfaces (APIs), and data models that are distributed to better align with temporally and spatially changing traffic loads.

Much of today's architecture is built upon components of a legacy cellular design. 1G networks focused on providing analog circuit-switched voice services over a wireless interface. 2G added digital voice capabilities and increased capacity. Mobile packet-switched data capabilities were integrated into the 2.5G/3G core network with packet data protocol (PDP) context support and added security features. The 4G network transitioned to an all-Internet protocol architecture for both voice and data and moved mobility management to the core, away from the distributed towers and controllers. In order to address 5G requirements, the following inefficiencies in the 4G network must be resolved:

- Mobility tracking is still based on legacy circuit-switched voice paradigms,
- Architecture is rigid and hierarchical,
- Header overhead is applied on every packet,
- Packets are not routed using the shortest path, and
- Excessive signaling occurs in high-density IoT networks due to connection-oriented data.

Several software-defined technology enablers integrated into the evolving 5G core will provide flexibility in networking and mobility, context-aware routing, and wireless backhaul/access integration. These technologies include software-defined radios (SDR), cloud computing NFV, and software-defined networking (SDN).

The significant advances in terms of cost and efficiency have piqued the interest of mobile network operators, and 5G is the first chance to explore the ways in which virtualization may benefit the mobile telephony world. Seeking to achieve similar gains in efficiency, the 5G research community is examining ways in which these same concepts of centralization and virtualization can provide benefit to their own networks. In general, the areas with the most potential benefit from virtualization are consolidating expensive base station equipment (Cloud RAN) and increasing the flexibility of deployments through rapid reconfiguration of the network.

This dynamic architecture also requires a comprehensive security architecture, which is scalable to meet the end-to-end protection of the user and control data as it traverses the various network nodes. Virtualizing components with software-configurable control opens up additional access vectors that must be considered and properly secured. The next two subsections will describe these new technologies and possible security issues.

3.1.1 Network function virtualization (NFV)

The core concept of NFV is to virtualize network functions that are traditionally hardware based into a virtual machine hosted in a cloud environment. NFV enables network providers to move towards a decentralized network, pushing core functions towards the RAN or network access edge, and virtualizing those functions on cloud-based servers. NFV is currently geared towards traditional IT network functions such as firewalls, domain name servers (DNS), deep packet inspection (DPI), and security gateways. However, core network functionality can also be hosted in virtual machines on high-speed, general purpose, commercial off-the-shelf (COTS) servers in a cloud computing environment. These functions include:

- Radio network controller (RNC),
- Mobility management entity (MME),
- Serving general packet radio service support node (SGSN), and
- Internet protocol multimedia subsystem (IMS).

Cloud and network operating systems such as OpenStack and OpenDaylight must meet carrier-grade deployment requirements in order to make NFV a secure and viable solution for future 5G networks. The primary concerns for a 5G network include availability, security, system performance, and network management. From the availability and reliability angle, there should be no single point of failure, and automatic detection and mitigation of faults becomes a mandatory requirement. Security of the cloud

infrastructure also becomes an issue. Process isolation between the various network functions must be stringent, with hardened protocols for authentication and validity of the command and control interfaces. On the performance side, the benefits of virtualization are moot if the speed at which the network operates must be reduced. Any virtualized environment must operate at speeds equal to its hardware predecessors. The final consideration is in the management overhead of the system as a whole. Scheduling, orchestration, and automated deployment of the virtual appliances must be simple and cost-effective to ensure that the benefits of virtualization are not outweighed by the overhead costs involved with maintaining the network.

Securing network functions based on virtual machines hosted in a cloud environment inherit the same security vulnerabilities currently encountered in cloud computing platforms, both publicly and privately hosted. Problems that need to be addressed include hypervisor security; isolating virtual machines (VMs) from buffer overflows, memory leaks, and interrupts; validating and authenticating users; and integrity checking virtualized network function (VNF) images. A quick scan of one of the most common open-source cloud operating systems (OpenStack) Common Vulnerabilities and Exposure (CVE) list indicates approximately 25 new vulnerabilities discovered in the past year (2016–2017) [11]. Many of these vulnerabilities are rated as a low probability of causing DoS and do not require authentication to carry out the attack. Two vulnerabilities with low to medium complexity allow for code execution on specific services within OpenStack.

The European Telecommunications Standards Institute (ETSI) NFV industry standards body has identified key areas of potential concern, summarized in the list below, which need to be addressed in the future 5G security architecture plan [12]:

- Topology validation and enforcement;
- Availability of management support infrastructure;
- Secured boot;
- Secure crash;
- Performance isolation;
- User/tenant authentication, authorization, and accounting;
- Authenticated time service;
- Private keys within cloned images;
- Backdoors via virtualized test and monitoring functions; and
- Multi-administrator isolation.

3.1.2 Software-defined networks

In order to support the growing demands of 5G, such as diverse mobile traffic patterns, massive capacity for the volumes of interconnected devices, application specific routing, complex connectivity, all across a heterogeneous infrastructure, a new network paradigm is required—SDN. SDN is an agile networking architecture based on open protocols, which decouple the network control functions such as routing and filtering from the packet processing hardware. This enables the network control to be software configurable, abstracting the underlying infrastructure for applications and network services. The abstractions define the components, the functions they provide, and the protocol to manage the forwarding plane. The network is abstracted into multiple, decoupled layers: the data plane, controller plane, and application plane.

SDN does not currently support mobility management. A new technology paradigm, software-defined wireless networking (SDWN), would provide the flexible control of radio resource, mobility, and routing management. As SDWN matures, it will enable new networking capabilities within 5G such as multi-homing, dynamic channel configuration, and session continuity.

SDN in 5G requires a strategy for securing the control plane traffic. The security architecture will need to protect the applications, data, and infrastructure from vulnerabilities introduced by the virtualization technology. The southbound APIs and protocols are primarily based on dedicated operating systems such as OpenFlow, Open vSwitch, OpenDaylight, and Open Network Operating System (ONOS) to name a few. Each of these operating systems is not invulnerable and they track new CVEs based on authentication, confidentiality, integrity, and availability attacks. Issues such as topology spoofing, SQLite memory leaks, authentication bypass, and MiTM attacks during key exchanges need to be resolved as the technology matures.

3.2 MMIMO- and mmWave-based air interface

Two new physical layer technologies are gaining massive interest from both the 3GPP and IEEE for use in 5G standards. The first is MMIMO, an antenna array technology that utilizes a very large number of antenna elements in a MIMO array, introducing the ability to provide a very large diversity gain to a large number of receivers. At traditional communications frequencies, such arrays would normally be far too large to be mounted on towers or the sides of buildings. However, a second emerging technology, mmWave, will enable the reduction of array size to a manageable dimension.

While both MMIMO and mmWave have the potential to be game-changers for the proposed 5G technologies, they introduce very little in the way of potential security concerns. Due to its propagation characteristics, a signal in the mmWave band is essentially a line-of-sight (LOS) transmission only. This means that antennas located indoors or in dense urban environments will not leak signals outside of their area of operation. A short-range LOS link provides a degree of assurance that any potential eavesdroppers will need to be very near the transmitter or receiver in order to overhear the conversation. This is further magnified if the MMIMO array utilizes smart antenna, or beamforming, techniques to narrow the directional beam towards the user, reducing the wireless channel to, effectively, a wired channel that does not need to be deconflicted through inefficient multiple access schemes. While these technologies can help prevent eavesdropping, they may introduce vulnerabilities of their own. For example, a sophisticated attacker may be able to use these tight beams against the system to reveal user locations or transmit interfering noise in mmWave bands with extremely small and stealthy jammers. In the first case, careful observation of known data fields may allow an attacker to reverse engineer the precoding weights applied to the transmitted data. Depending on the scheme used for precoding, there is a small chance that an attacker may be able to use that information to determine the location of the user, effectively trading data confidentiality for location confidentiality [13]. In the second case, the same benefits that make mmWave bands attractive for MMIMO makes them attractive for covert jamming: The small antenna size and relatively low power make it possible to easily design a pocket size jamming device that can deny communications to a large area.

3.3 Increased throughput and decreased latency

DPI enables operators to track the content of packets going through the network with the intent of ensuring security and QoS for individual users. DPI can be used to both manage the data networks globally and optimize individual user traffic by examining the full packet contents as opposed to just the IP headers. The most common use of DPI is to determine the type of application within a connection stream, such as HTTP, mail, streaming video, or peer-to-peer traffic with the intent of ensuring policy enforcement and charging rules.

Accurately identifying the application payload in packets is also important to ensuring that traffic is routed properly, charged appropriately, and malicious activity is detected. Due to the fact that new applications are introduced at a phenomenal rate, it can be difficult to maintain filters to sort the traffic and possibly block unwanted activity. Certain applications may also intentionally disguise their identities to bypass firewalls and other devices performing DPI. A database must be maintained to address the dynamic application signatures, and the system must be intelligent enough to quickly adapt to new applications without reducing itself to checking packets against a large number of individual filters.

For future 5G systems, the implementation becomes more complex due to a combination of rapidly evolving applications, increasing packet data rates, and strict low latency requirements. Even now, a DPI system has only nanoseconds to inspect a packet, modify it if necessary, and send it on to the next node without incurring buffer overflows. In order to keep up with the tens of Gbps throughput on current technology, multicore and multithreaded processors are required, making it difficult to virtualize the functions in this case. As previously discussed in the NFV portion of this article, forcing a hardware-only implementation will greatly increase the cost of widespread implementation throughout a 5G network.

A 2015 Gartner study on carrier-class network firewalls (CCNFWs) noted that communications service

providers implementing firewalls and DPIs will struggle with handling the demand from increased traffic from the new 5G paradigms such as IoT and increased heterogeneous mobility. These new services will also increase number of threats to the network. Product managers of CCNFWs need to develop roadmaps that include better integration with SDN and NFV environments, and better threat protection to meet future needs of communication service providers (CSP) [14].

4. Potential solutions

While many of the security problems facing a potential 5G standard seem very daunting, there are a number of potential optimizations and general tacks that can be taken when designing the overall system security. This section presents suggestions and potential implementations of security paradigms that support the various problems and issues outlined in the previous sections.

4.1 Multilevel security paradigms

As previously discussed, one of the greatest difficulties in building a 5G standard is going to be the integration of orthogonal technical requirements from the different use cases. While this task is daunting, it may be possible to utilize multilevel security paradigms to alleviate some of the more complex portions of the reconciliation.

Multilevel security paradigms work by treating different classes of data with different levels of required security. For example, data may be protected based on many different factors: its origination point, the type of content being carried, or the level of trust the network has for a particular user. A few examples of this type of protection could include video streaming, redundant sensing data, or self-driving car data. The first example demonstrates how data confidentiality may not be necessary for all connection types. A user may be broadcasting a video from their phone—streamed to the Internet for a public performance. If the user has marked this video as public, there is little need to encrypt the data for confidentiality: A would-be attacker sniffing the network would be able to access the data just as easily by subscribing to the user's multicast IPv6 feed. The second example provides a use case where authentication may be unnecessary altogether. For redundant sensing data, a few bad actors will not be able to appreciably affect the sensed parameter without taking control of a large number of network nodes—authentication for every data point is not necessary. Finally, self-driving cars provide a scenario in which disregarding the availability requirement may be beneficial. In a theoretical network of cars running down the highway, conditions are changing rapidly—it makes no sense to ensure that data which may be several seconds old eventually gets delivered to the intended vehicle; by then the data is stale and useless. If the network had instead focused on transferring only relevant data, the cars may have been able to exchange data at a higher rate.

4.2 Reduction in data monitoring

The number-one contributing factor to preventing attacks within a network is good awareness of the types of traffic that are flowing across the links. While DPI and content analysis systems are very good at rooting out bad actors, they are also resource intensive and introduce significant delays into networks, even at today's bandwidths. With the bandwidth explosion on the horizon for 5G, it is clear that DPI and content analysis will not be possible on every link running across the network. The number-one task for increasing the operator's awareness of the network will involve being smarter about what gets inspected. Through the use of machine learning and data analytics, network security within the packet core will become a question of which packets to inspect that ensure the highest probability of catching attacks and stopping them before they become a problem. In addition, the idea of utilizing cross layer information to drive that decision-making process will become increasingly important. For example, the activity of an adversary spoofing another user's connections may fall within normal ranges at the network level, but may be very odd at the access layer. While this behavior would not be noticed utilizing traditional security monitoring tools, combining the information about the odd access behavior with analysis of the network layer connections would reveal an immediate threat.

4.3 Distributed trust models

As the size of the 5G network grows and the connection paradigms become less and less fixed, the trust model for mobile telephony is going to have to move

away from its current "central authority" model. With the extremely large number of MTC devices predicted to come online and the ability for D2D communications to happen at arbitrary points in the network, there must be a method for establishing trust between both pairs of arbitrary users and an arbitrary user and the edge equipment in a network. There are several candidate technologies for distributed trust that have been under research for decades [15–17]. Any of these models, coupled with the current central trust model, will allow for a reduction in network congestion and allow safe out-of-band operation for edge devices. When a secure system is not possible using a strictly cryptographic mechanism, alternative methods of authentication can be explored including caching the credentials of users likely to be in the area, utilizing a distributed block-chain system similar to the Bitcoin electronic currency, or a modified version of the Distributed Authentication Security Service (DASS) [18].

4.4 Securing virtualized networks

There are some added security benefits to NFV, including the ability to virtualize security services on demand, such as malware protection, cloud-based URL filtering, and web application security. The features can be easily customized per customer use case and delivered via a cloud-based portal. A centralized security-as-a-service model simplifies the management of a common security policy across diverse virtualized network functions. This was not possible for application-specific hardware solutions since each router or switch required that the security services were tailored to the hardware framework, leaving the possibility of inconsistent security policies throughout the network.

Securing NFV will be a recursive process as compute, storage, network, and orchestration resources are optimized. Several mobile telecom solution providers are offering cloud-based NFV platforms with a policy-driven approach to orchestration, security zoning, and workload placement. The security policy can be specified using the standard OASIS Topology and Orchestration Specification for Cloud Applications (TOSCA) language. The built-in automation capabilities can proactively and reactively remediate security problems as they arise.

In order to increase the reliability of 5G networks, the security implementation of the SDN controller and management needs to be improved. The controller is the centralized decision point for network traffic and therefore must be tightly controlled. If the SDN controller goes down due to a distributed DoS (DDoS) attack, network availability is nonexistent. Increased protection at the control layer should minimize downtime due to attacks. Communications throughout the network must be protected with an adequate trust model. The SDN controller, the applications loaded on it, and the devices it manages need to be trusted entities with integrity protection. The addition of a robust policy framework applied to the architecture will supply a system of checks and balances to ensure the controllers are functioning correctly. And finally, when an incident does occur, forensics and remediation tools need to be in place to quickly and accurately determine the cause of failure, recover, potentially report on the event, and then protect against future attacks of similar signature.

While none of these solutions provides a clear avenue towards total security within the network, they do provide hope that the reconciliation of security paradigms in 5G networks is not insurmountable.

5. Conclusions

There is no doubt that numerous 4.xG-version architectures will be studied as networks are upgraded, but ultimately there should be a single interoperable 5G standard that will cover all mobile communications from the slowest to the fastest. The current direction for higher speed is focusing on 10 Gbps connections over the air—enough to allow hospitals to send real-time imagery over the Internet to remote consultants. Additionally, the architecture will enable both low latency and the ability to connect 10 billion devices around the world. At this point it is clear that current security paradigms in the mobile telephony industry are going to be sorely inadequate for the 5G cellular systems deployed to meet the ITU 2020 requirements. The ITU requirements introduce a daunting combination of increased bandwidth, competing use cases, and ultradense device connections. The reality of providing service for all user connections under these constraints will quickly outpace the current technology used for traffic monitoring, user authentication,

and privacy guarantees. While the problem is not insurmountable, there will likely be a number of difficult technical issues, each with interdependent constraints, which will require large number of compromises to meet the ITU's requirements.

In general, a decentralization of the security paradigm will be required. It is simply a matter of fact that current and projected computing power will not be sufficient to scan all of the bandwidth across 5G networks with sufficient speed to meet the latency requirements. New models will need to be developed that will allow for the end-to-end security for user data without a centralized monitoring and control system. Adding to the decentralization theme are the new use cases such as edge computing, D2D communications, and V2V networks; all of which need to operate without a central authorization authority. These use cases simply cannot exist without authentication methods that do not rely on the current central authority model.

In the end, it seems likely that a combination of several solutions will be required in order to meet both security and performance constraints in 5G networks. Standards organizations advocating a "phased roll out" model for their 5G protocols must remain aware of all of these existing constraints during each phase of development to ensure that initial phases do not introduce unnecessary difficulties for the later phases. An initial focus on the eMBB use case may provide the best return on investment for industry and spur 5G adoption, but it is not the most difficult problem from a security standpoint; great care must be taken from the outset to ensure that security meets the constraints of all proposed use cases.

References

[1] ITU-R. "IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond." Recommendation ITU-R M.2083-0. September 2015. Available at: http://www.itu.int/rec/R-REC-M.2083.

[2] *IEEE International 5G Summit;* 2015 May 26; Princeton, NJ. Available at: http://www.5gsummit.org/index.html.

[3] *3GPP 5G RAN Workshop;* 2015 Sep 18-19; Phoenix, AZ. Available at http://www.3gpp.org/news-events/3gpp-news/1734-ran_5g.

[4] Roessler A, Schlienz J, Merkel S, Kottkamp M. "LTE-Advanced (3GPP Rel.12) technology introduction." Rohde & Schwarz. 8 April 2015. Report No. 1MA252. Available at: https://www.rohde-schwarz.com/us/applications/lte-advanced-3gpp-rel.12-technology-introduction-white-paper-white-paper_230854-108294.html.

[5] Bilogrevic I, Jadliwala M, Hubaux J. "Security issues in next generation mobile networks: LTE and femtocells." *2nd International Femtocell Workshop;* 2010 Jun 21; Luton, UK. EPFL-POSTER-149153.

[6] Aazam M, Hung P, Huh E. "Smart gateway based communication for Cloud of Things." *2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP);* 2014 Apr 21–24; Singapore. doi: 10.1109/ISSNIP.2014.6827673.

[7] Cirani S, Ferrari G, Iotti N, Picone M. "The IoT hub: A fog node for seamless management of heterogeneous connected smart objects." *12th Annual IEEE Conference on Sensing, Communication, and Networking—Workshops (SECON Workshops);* 2015 Jun 22–25; Seattle, WA. doi: 10.1109/SECONW.2015.7328145.

[8] Jones J. "Edge computing: The cloud, the fog and the edge." *SolidRun.* 2017 April 23. Available at: https://www.solid-run.com/2017/04/23/edge-computing-cloud-fog-edge/. [This reference and corresponding figure was added after original publication of article in SPIE.]

[9] Zarif-Kadir, N. "Top 5 reasons to choose cloud computing." WebSan Solutions Inc. 2015 Jan 6. Available at: http://www.websan.com/blog/item/560-top-5-reasons-to-choose-cloud-computing.

[10] Nair K, Kulkarni J, Warde M, Dave Z, Rawalgaonkar V, Gore G, Joshi J. "Optimizing power consumption in IoT based wireless sensor networks using Bluetooth Low Energy." *2015 International Conference on Green Computing and Internet of Things (ICGCIoT);* 2015 Oct 8–10; Delhi, India. doi: 10.1109/ICGCIoT.2015.7380533.

[11] MITRE Corporation. CVE Details. Openstack: Security Vulnerabilities. 2017 June 29. Available at: https://www.cvedetails.com/vulnerability-list/vendor_id-11727/Openstack.html. [This reference and corresponding information was added after original publication of article in SPIE.]

[12] Briscoe B. "Network functions virtualization (NFV); NFV security; problem statement." European Telecommunications Standards Institute. October 2014. Report No. GS NFV-SEC 001 v1.1.1.

[13] Wong K T. "Blind beamforming/geolocation for wideband-FFHs with unknown hop-sequences." *IEEE Transactions on Aerospace and Electronic Systems.* 2001;37(1):65–76. doi: 10.1109/7.913668.

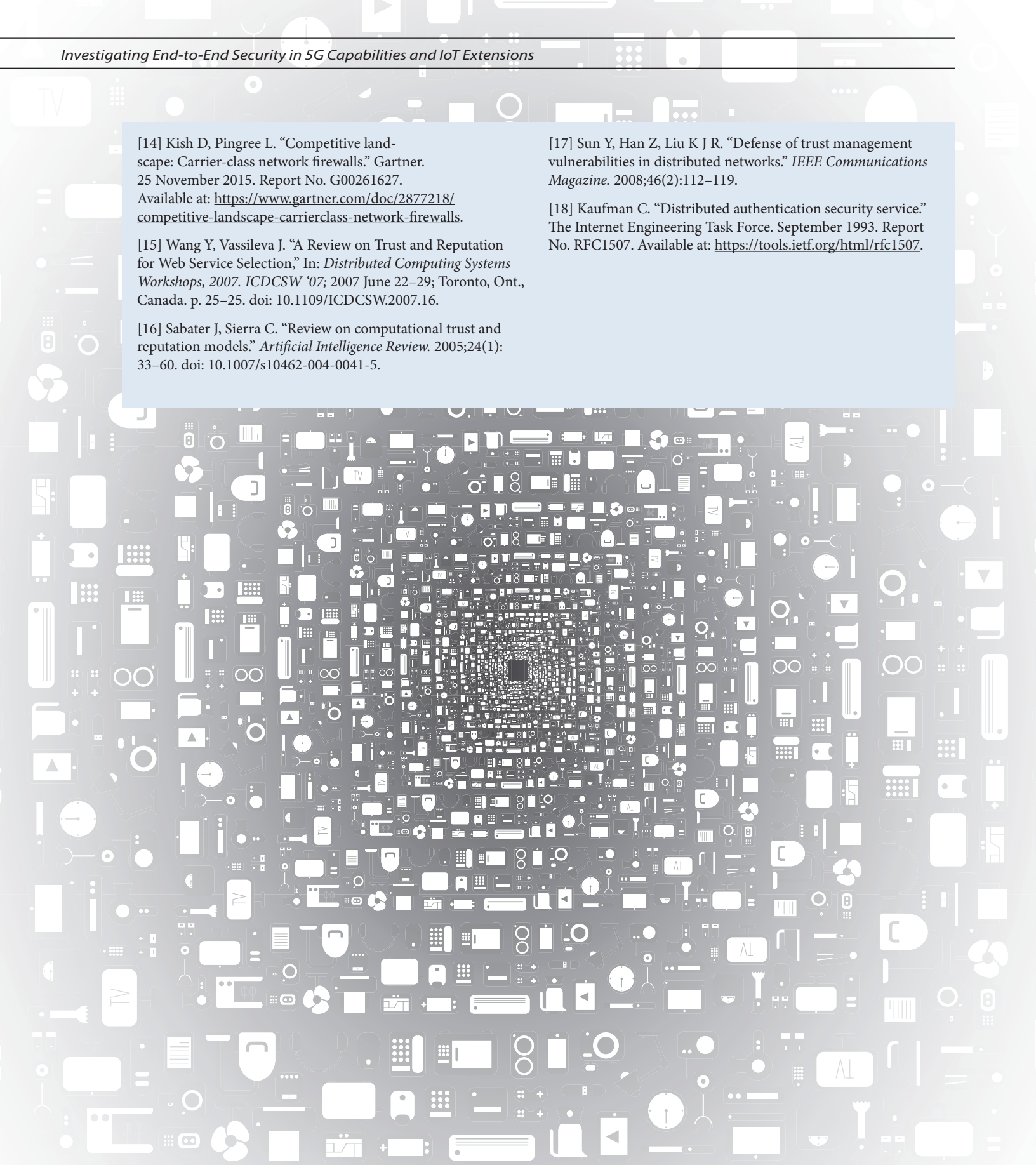
[14] Kish D, Pingree L. "Competitive landscape: Carrier-class network firewalls." Gartner. 25 November 2015. Report No. G00261627. Available at: https://www.gartner.com/doc/2877218/competitive-landscape-carrierclass-network-firewalls.

[15] Wang Y, Vassileva J. "A Review on Trust and Reputation for Web Service Selection," In: *Distributed Computing Systems Workshops, 2007. ICDCSW '07;* 2007 June 22–29; Toronto, Ont., Canada. p. 25–25. doi: 10.1109/ICDCSW.2007.16.

[16] Sabater J, Sierra C. "Review on computational trust and reputation models." *Artificial Intelligence Review.* 2005;24(1): 33–60. doi: 10.1007/s10462-004-0041-5.

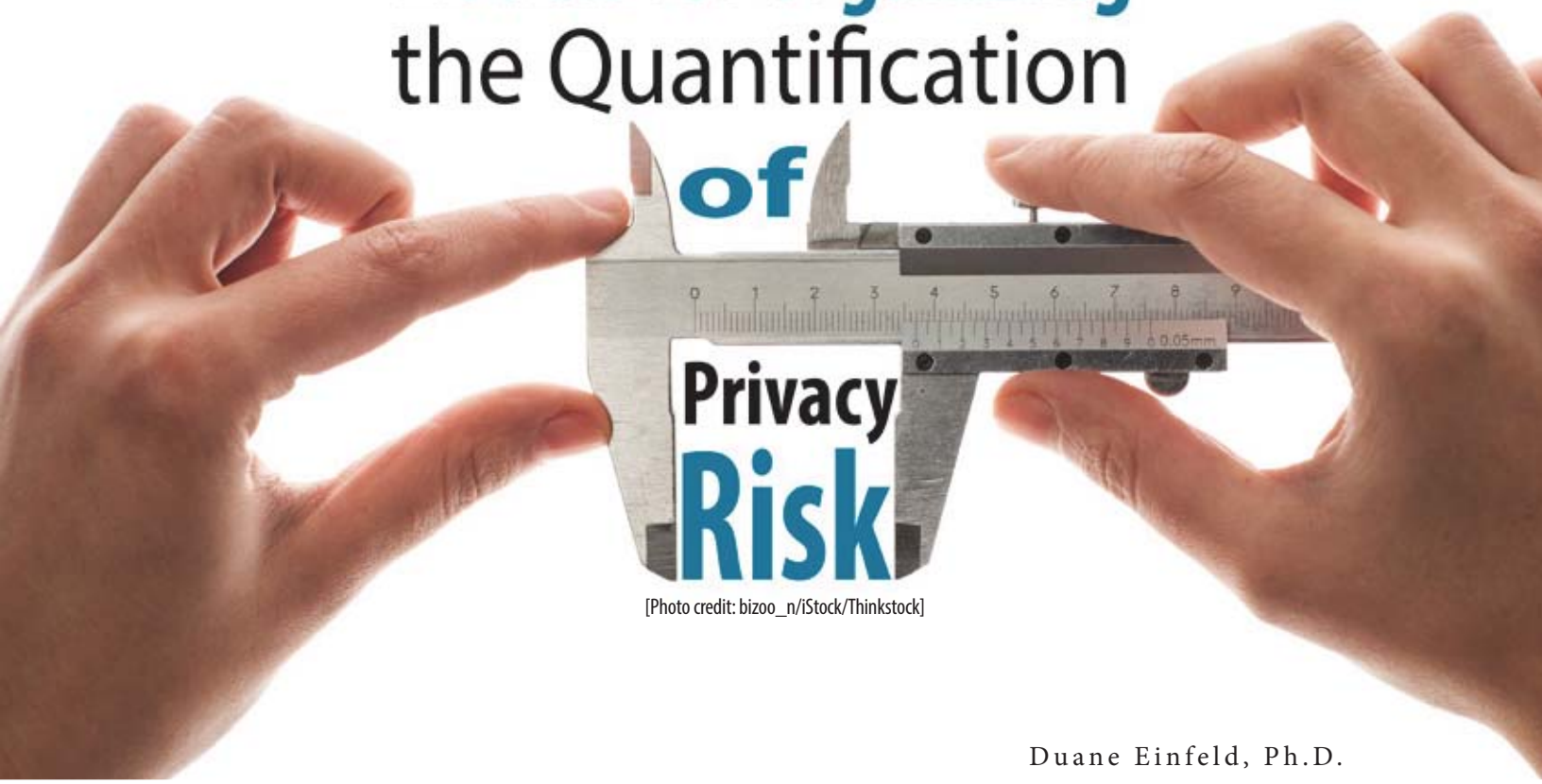
[17] Sun Y, Han Z, Liu K J R. "Defense of trust management vulnerabilities in distributed networks." *IEEE Communications Magazine.* 2008;46(2):112–119.

[18] Kaufman C. "Distributed authentication security service." The Internet Engineering Task Force. September 1993. Report No. RFC1507. Available at: https://tools.ietf.org/html/rfc1507.



[Photo credit: Kroljaaa/iStock/Thinkstock]

Models for Organizing the Quantification of Privacy Risk



[Photo credit: bizoo_n/iStock/Thinkstock]

Duane Einfeld, Ph.D.

The wealth of personal data on the Internet, which is held by both corporations and governments, has raised public concern about information privacy. This has led to research about how to protect individuals' and groups' privacy both online and in other settings. The author, recently involved in such science of privacy research, considered how to quantify privacy risk of data, and here he presents two rudimentary models that have come from the effort. The first is a graphical way to visualize the interactions among data relationships and data uses that affect privacy risk. The second model, a matrix that lists privacy categories and personal loss types, is a tool intended for estimating levels of privacy risk or levels of potential loss resulting from privacy compromise in a given set of personal information data. Although the mathematics of the models, especially the first one, is not fully formed, their rough reflections of notions of privacy may provide a beginning for further privacy research and eventual application.

People, corporations, and government entities use the Internet for everything from social media, online shopping, and webmail to marketing research, employment data, and health data. Consequently, the Internet is using much data about many people in many ways, and this raises privacy concerns. There may be good reasons for personal data to reside in corporate or government databases, but this introduces risk that the data could be shared with those who do not have good reasons to have the data. Even those who have good reasons to access others' personal data might combine it and mine it to discover information that was not intended to be shared. To better understand how personal information should be protected, it would be useful to quantify the risks involved with handling personal information.

My objective here is to suggest one or more ways to organize thoughts about privacy and how to quantify privacy risk.

Here we will refer to such privacy-related data as personal information (PI). Note our attention here is on PI in general, not strictly personally identifiable information (PII). PII is information that by itself is enough to identify a particular individual, such as a Social Security number. However, privacy concern extends beyond PII to include data that when used in combination with other data may be enough to identify an individual. It also includes information that might not be unique to the individual, but the disclosure of which, if associated with the individual, would be perceived to be personally injurious to the individual.

Measures to reduce risk of PI leaks, and hence of privacy loss, need to involve several components. One component is laws that restrict the sharing of personal information. Another is secure databases and procedures to prevent unauthorized access to sensitive personal data. However, data security has costs, so custodians of data need to exercise judgment about which PI should be kept more secure and which PI may be kept less secure to deal with threats from both unauthorized and authorized persons.

All the ways of handling PI create risks to privacy. To judge how well a set of security measures are protecting privacy of PI, it would be beneficial to quantify privacy risks based on how accessible various types of PI are and how likely they are to be combined for inferring other PI. Here, risk includes not only the likelihood of intentional or unintentional disclosure of the privacy-sensitive data, but also the (tangible or intangible) loss, or impact, that could result from such disclosure.

That is, the risk we are concerned with quantifying is not merely the level of *security* of personal information in a database; it is the level of personal injury of the *privacy compromise* that could result from the types of personal information in the database and how it is stored or handled.

Steps for quantifying privacy risk

Attempting to quantify privacy risk involves several elements; I suggest a breakdown of roughly four overlapping steps:

1. Make philosophical assumptions about the possibility of quantifying privacy,
2. Identify categories of data for privacy rating,
3. Assign the privacy ratings, and
4. Identify effects of data uses on privacy ratings.

Among the philosophical issues of privacy (in step 1) is the question: Given the diverse, subjective opinions of how private personal information is, is it at all feasible to quantify it on a single fixed scale for everyone, or even on an adjustable scale (adjustable based on a set of factors that differ for different individuals)? Privacy, a social concept, is less precise than security. But there is commonality among opinions about privacy; for example, people view Social Security numbers and personal health information as very privacy sensitive, while viewing purchasing habits as less privacy sensitive [1]. Also, laws are made about privacy. Thus, we may assume there is enough agreement in people's thinking about privacy sensitivity and risk that we can start creating a model to quantify them.

Identifying categories of privacy data (step 2), which we will call *personal data types,* includes the challenge of creating enough categories to be meaningful (enough distinctions) and few enough categories to be manageable (few enough to catalog). Example types might be "name," "address," and "Social Security number" (of an individual), but there may be reasons to define broader categories.

Assigning privacy values (step 3) requires us to decide which kind of value we mean: *privacy sensitivity* (people's sense of which data types are more private than others), *likelihood of loss, magnitude of potential loss,* or *level of risk,* for example. Values might be assigned to individual categories or to categories in combination. They might be assigned by experts, by a crowd-sourced survey of a population, by data from social networking sites, or a by combination of these.

Finally, data is not meant to be entirely static; it gets used. It thus makes sense (step 4) to define categories of uses and then to try to characterize the effects that particular analytic uses of data may have on privacy by use category. Uses include loading data into a database, combining data, filtering data, or deleting data, for instance, any of which could affect data privacy risk. Encrypting or obfuscating data could also have effects.

Below I present two models to suggest how to begin to manage these steps. The models are partial; neither includes all four steps, and they are not (yet) combined into a single model. One model—a set of graphs—assumes relevant categories of PI exist; it focuses on relationships of PI categories and uses of data and the influence of both on privacy. The other model—a matrix tool—suggests categories of personal information and loss to have in mind when subjectively assessing levels of privacy risk or potential personal loss for a set of data.

Graphical model for quantifying privacy risk

As asserted above, assessing privacy risk of personal information requires categorizing the data into personal data types and then assessing risk by category. People tend to view different personal information types as having different levels of privacy sensitivity.

However, here we will take the view that privacy sensitivity does not make sense for personal data types *individually.* It should apply only to personal data types *in combination* (i.e., when it is known that the types are associated with each other). For example, a password by itself is just a series of characters, but if it is known to be a password and known to be for a specified online bank account of a specified person, then another person with knowledge of this combination of information can become a threat to the account and its owner.

Consequently, the first model proposed here for our assessment of privacy risk depicts graphically the relationships (i.e., associations) of personal data types in a database—which is to say, the relationships that are known to us. Also, because analyzing—or using—data changes how much we know about relationships within the data, the graphical model of personal data type relationships is expanded to incorporate data uses.

Correspondingly, the model consists of two types of graphs: the data relationship graph and the data use graph.

The data relationship graph

The data relationship graph (DRG, see figure 1) is intended for estimating privacy risks of data based on data relationships that exist at a given time. It

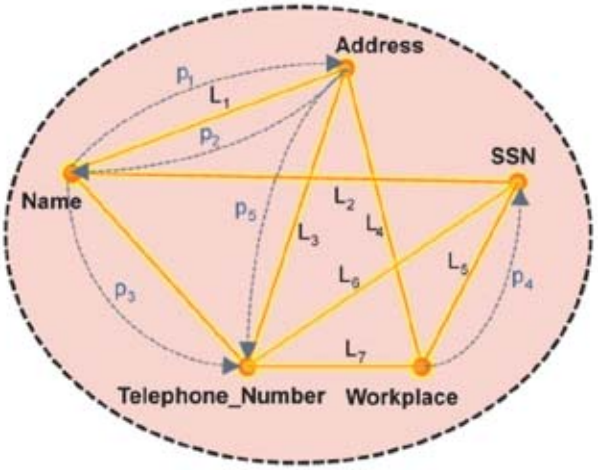


FIGURE 1. Data Relationship Graph. Nodes represent data types/fields. Edges represent privacy loss level (negative impact) *L*. Arrows represent probability (likelihood, ease) *p* of combining two data types, conditioned on knowing one of the two. (For ease of reading, the graph shown here is not drawn complete; an edge and two arrows could be drawn between each pair of nodes.)

represents types of data and relationships among them, along with levels of possible privacy loss that could result from associating data of two or more types. (We will refer to loss rather than impact and say that positive values indicate losing privacy, and negative values represent gaining privacy.)

Each node of a DRG is a personal data type, such as name, address, or Social Security number. The relationship values in a DRG are values describing the association of two or more data types as known by a database. One kind of relationship value is the potential loss (*L*) of possessing knowledge of the values of two (or more) personal data types for an individual, such as knowing both a person's name and address, L_1 in the figure. This is represented by an edge between the two vertices having the value L_1 assigned to it.

The other kind of data relationship value is the conditional probability (*p*) of determining the value of one data type given knowledge of another type for the same individual, such as the probability of knowing a person's address given that one knows his or her name (p_1 in the figure), or vice versa (p_2 in the figure). These are represented by arrows in the DRG. (It is not strictly necessary that the model use genuine probabilities; perhaps a different measure of the ease of

associating data or the time to associate data would be beneficial, maybe even with values chosen subjectively. Nevertheless, for now we think in terms of actual probabilities.)

To provide a simplistic example using conditional probabilities, suppose we have data in the form of two lists. One lists names and addresses, in which of all names listed, only fraction p_1 of them have addresses listed. The other is a list of addresses (including all those from the first list) and telephone numbers, in which of all addresses listed, fraction p_5 of them have telephone numbers listed. We are able to estimate from this, assuming the probabilities are independent, that given a listed name of a person, there will be probability $p_1 \times p_5$ of determining the person's telephone number using these two lists.

Having defined the data relationship graph, we then associate a level of risk with the graph, or with subgraphs of it. Risk involves not only levels of potential loss, but also the likelihood of experiencing such losses. The numerical value of risk (R) is commonly defined as the sum of the probability of loss $p(e)$ of event 'e' times the level of loss $L(e)$ of event 'e,' over all events $e \in E$, where E is the set of all possible loss events. That is:

$$R = \sum_{e \in E} p(e)L(e).$$

For the examples we have just discussed, we can say that given a name on the list of names and addresses, the risk of associating the name with an address is $R = p_1 L_1$.

Knowing the risk values (R) for pairs of data types for an individual, it would be useful to estimate the risk for larger combinations of data types. For example, if individuals' privacy risk for starting with their names and associating their addresses is $R_1 = p_1 L_1$, and the privacy risk for starting with their addresses and finding their telephone numbers is $R_2 = p_5 L_3$, then the privacy risk for starting with name and finding the other two must be at least the maximum of $p_1 L_1$ and $p_1 p_5 L_3$ (the second expression uses the probability of finding a telephone number based on a name, $p_1 p_5$). This is one idea for handling risks in combination; likely others could be devised.

The data use graph

As noted above, data relationships may change as data is used. That is, by applying data analytics—such as sorting, correlating, and filtering operations—associations of data types may become stronger or weaker, or appear or disappear. If so, we say there is a change in the state of the data in the database ('state' meaning the set of data and relationship values at a particular time), an effect that shows up as a change in the potential loss and probability values assigned to edges and arrows of the data relationship graph.

Two clarifications of this theoretical model are in order: 1) Instead of saying that state changes create or delete data types and relationships, we will say the model has all data types and relationships existing at the outset and remaining, with only the probability and loss values changing. (Figure 2, below, will not show that, though.) 2) We are using the word "database" to refer to all the data we have access to, whether it is in just one computer system or several. If data is stored in one computer and exported to another to be analyzed, we still consider data in both as being in the database. We are modeling the information, not the system.

The data use graph (figure 2) represents how analytic uses of data effect state changes. Each node represents a state of the data, shown as a data relationship graph. Arrows representing uses, shown extending from one node (one DRG) to another, indicate how each use potentially changes data relationship values from one state into another.

The nodes are numbered (in the figure, $j = 1, 2, 3,$ or 4). Node number j (i.e., state j) is labeled R_j (R is short for DRG). Each arrow (each use) is labeled $u_{R_j R_k}$, meaning the use that changes state R_j (at the tail) into state R_k (at the head). The main changes from one state to another are the probabilities. If $\boldsymbol{P}_j$ is the list of conditional probabilities in state R_j, and $\boldsymbol{P}_k$ is the list of conditional probabilities in state R_k, then we write:

$$\boldsymbol{P}_k = u_{R_j,R_k}(\boldsymbol{P}_j).$$

Perhaps with sufficient understanding of the nature of privacy it would be possible to identify mathematical functions u that approximately reflect the effects of particular uses. Maybe a correlation use increases the probability of associating two personal data types; maybe a filtering use decreases the probability. It merits further investigation.

This is the extent of the research done on the two-graph model to date.

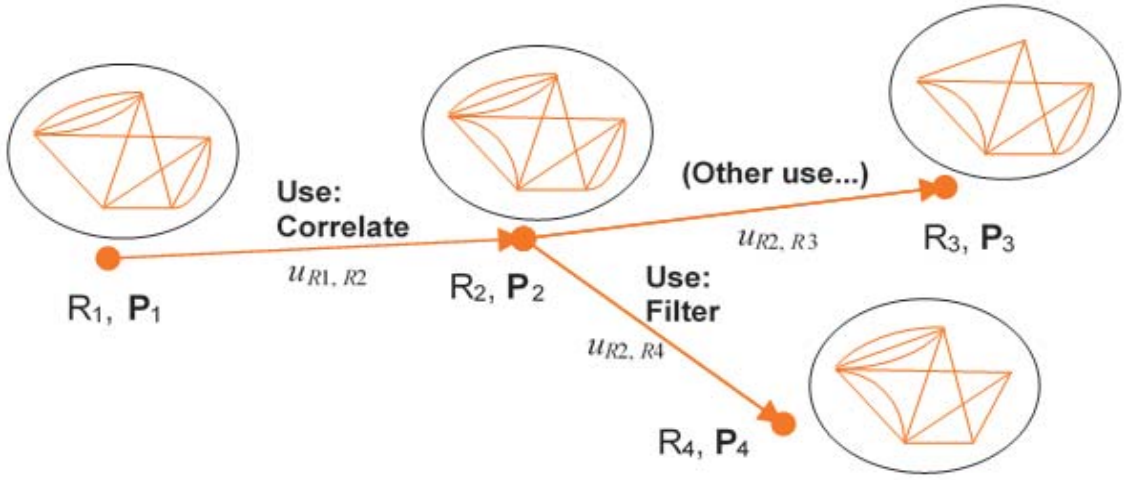


FIGURE 2. The Data Use Graph. Each node corresponds to a state of the data at a given time, and thus each has an associated data relationship graph. Arrows between nodes indicate analytic uses performed on the data that change the state of the data. (The DRGs shown are merely suggestive of states having stronger or weaker known relationships, indicated by more or fewer arrows between vertices.)

Matrix tool for quantifying privacy sensitivity

The two-graph model described above, representing data relationships and uses, is a rudimentary, incomplete approach to the problem of quantifying privacy risk. It does not yet define how to assign loss values and probabilities (or probability-like values) to edges and arrows. Nor does it offer a definitive formula for combining the graph components and weights into an overall measure of privacy risk. It is hoped further research could inform these choices.

Setting aside the question of how to assign probabilities, we will consider how to assign privacy risk or loss values, but not in conjunction with the two-graph model, though perhaps it could be incorporated into that model. Privacy sensitivity was mentioned above; we will interpret it ambiguously, to mean either a level of risk of, or else a potential loss from, privacy compromise.

As noted earlier, it is not obvious how to construct a scale representing privacy sensitivity for PI data types that would be useful for society collectively; it would be difficult even to reflect the attitudes of more than one person. Privacy sensitivity may be very subjective and thus hard to quantify. On the other hand, the need for privacy seems essential to being human, and people have at least a degree of common sentiment regarding what kinds of information are more private than others.

To reflect these sentiments, while still lacking an *objective* measure of privacy sensitivity, we will consider a scheme for assisting our *subjective* judgments of privacy sensitivity. Specifically, since we want to relate personal information to potential loss, it may be helpful to create a set of categories for each. The first set will be *categories of privacy* within which to assign loss values. The second set will be *types of personal loss* that could result from the compromising of personal information. It is conceivable these categorizations may refine our focus and so improve our ability to quantify potential loss.

Categories of privacy

To identify categories of privacy, we can start with categories drawn from the Fourth and Fifth Amendments of the Constitution of the United States, as follows (emphases added):

> **Amendment 4.** "The right of the people to be secure in their *persons, houses, papers,* and *effects,* against unreasonable searches and seizures, shall not be violated …"[2].

> **Amendment 5.** *"No person … shall be compelled* in any criminal case to *be a witness against himself,* nor be deprived of life, liberty, or property, without due process of law; nor shall private property be taken for public use, without just compensation"[3].

The Fourth Amendment's listing of "persons, houses, papers, and effects" protects what might be called privacy of an individual's ***body, personal space, personal records (including PII),*** and ***personal possessions.*** The Fifth Amendment's right not to testify against oneself suggests one's ***thoughts*** are to be respected as private. Research by others suggests the categories of ***personal relationship*** information and ***biographical*** information may also be useful. (Often, biographical information may be used as identity information, but here we are attempting to distinguish standard PII, such as a one's driver's license number, from life history information, such as the make and model of one's first car.)

Consequently, we have this representative list of privacy categories (or privacy information categories):

1. Thought life (including feelings, preferences, communication)
2. Personal relationships (family, friends)
3. Body information (including health)
4. Life history (biography, personal patterns)
5. Personal space (including home, website, e-mail account)
6. Personal economics (including property, business relationships)
7. Identity (forms of ID, passwords, online identity).

To help distinguish the categories and to make them more memorable, I have ordered them roughly according to a kind of personal distance from the individual. Establishing this order for the privacy categories is done only to facilitate understanding; it does not affect the privacy assessments we might base on the categories.

Types of personal loss

In the influential article "The Right to Privacy," Samuel D. Warren and Louis Brandeis identify the "right to privacy" as a "right to be let alone" [4]. To settle on that definition, Warren and Brandeis discuss it—and the associated mental suffering from the violation of that right, e.g., violation by the press, by a photographer, or by the possessor of a recording device—in comparison to several areas of law that are very similar. Among these areas are: defamation (slander and libel); copyrights; property rights; right to publish or prevent publication; breach of trust, confidence, or contract; liberty (freedom from restraint); freedom from injury, imprisonment, or malicious prosecution; and trade secrets.

This leads us to a potential list of types of loss: *anonymity, liberty, money, opportunity, personal peace or safety, property, reputation, solitude,* and *trust or confidence.* Recognizing that in the realm of information, one kind of personal information loss can lead to another, we add to this list the *loss of secrecy of other information.* To reduce the number of items, we combine a few categories, and the result is the following list:

1. Loss of liberty (things allowed to do)
2. Loss of trust or confidence
3. Loss of personal peace, safety, or solitude (including anonymity)
4. Loss of reputation
5. Economic loss
6. Loss of opportunity (things available to do)
7. Loss of secrecy of other information

The order of the types seems immaterial; I have attempted to order them according to a possible scale of importance, but this order could easily be debated. The list of loss types is similar to a "Catalog of Problems for Individuals" listed by the National Institute of Standards and Technology (NIST)[5].

Privacy sensitivity matrix

Having settled on a set of privacy categories and a set of loss types, we can use the two lists as row and column headings, respectively, in a matrix (table 1). Weights may be assigned to the privacy categories or to the loss types or to both to indicate importance for a particular context; here we limit ourselves to privacy category weights (see the first column). Scores may be entered into the matrix cells, and then the scores can be used collectively for assessing potential loss

TABLE 1. Privacy Sensitivity (Risk or Potential Loss) Matrix. Matrix entries are risk levels or potential loss levels [Low (1), Medium (2), High (3), or blank (0)].

		Personal Loss Type						
Privacy Category Weight: 0, 1, or 2	Privacy Category	Liberty	Trust/ confidence	Personal peace/ safety/ solitude	Reputation	Economic	Opportunity	Secrecy of other information
	Thoughts							
	Relationships							
	Body							
	Life history							
	Pers. space							
	Pers. econ.							
	Identity							

(or assessing risk) associated with the compromising of privacy. (Whether to assess risk or potential loss should of course be decided in advance of filling in the matrix.)

One possible way to use the matrix is the following: The user knows of data stored in a particular computer system or database and would like to roughly assess the privacy sensitivity of the stored data in terms of potential loss. Having the data in mind, the user may assign a weight (0, 1, or 2) in the left column to represent the significance of each privacy category. Then the user may work through the rows of privacy categories and for each cell within a row, state which level of potential loss is associated with the loss type for that column, for the data set in question. Perhaps the levels to be assigned will be Low (1), Medium (2), High (3), or none (blank or 0).

After the user fills in sufficiently many cells, each row weight is multiplied by every cell value in that row, and then all the resulting products are added together for the entire table. Low results of the calculation represent low loss; high results represent high loss.

A second possibility is to enter levels in the matrix cells explicitly as 'Low,' 'Medium,' and 'High,' then copy the corresponding row weight (0, 1, or 2) next to each word, and then add up those weights for each of the Lows, Mediums, and Highs in the table, to yield a set of three overall numbers (Lows, Mediums, Highs). This vector of three numbers may provide a better idea of how the losses are distributed than a single value would.

Future research

The research we have discussed leaves much room for follow-on work. For example, the risk associated with a data relationship graph could be given more specificity; the possible uses in the data use graphs might be identified by categories; and the privacy sensitivity matrix could be tested for usefulness on real systems and data—however, no such testing has yet been done. These ideas are offered for validation or further research, where the main issue to be tested is whether the eventual process or processes yield consistent results in line with intuitions about privacy sensitivity.

References

[1] Madden M. "Public perceptions of privacy and security in the post-Snowden era." Pew Research Center. 12 November 2014. Available at: http://www.pewInternet.org/2014/11/12/public-privacy-perceptions/.

[2] US Constitution. Amendment 4.

[3] US Constitution. Amendment 5.

[4] Warren S, Brandeis L. "The right to privacy." *Harvard Law Review.* 15 December 1890; 4(5): 193-220.

[5] Brooks S, Nadeau E, eds. "Privacy risk management for federal information systems (Draft)." May 2015. NISTIR Document 8062. Available at: http://www.csrc.nist.gov/publications/drafts/nistir-8062/nistir_8062_draft.pdf.

Device-to-device communication: LTE Direct

Staff Writer

LTE Direct is a device-to-device (D2D) communication technology that is expected to play a role in the development of fifth-generation (5G) wireless technology, supporting a larger and more diverse set of devices and applications. The appeal in using LTE Direct for 5G D2D communications is its ability to transmit over long distances at high data rates. This is particularly beneficial in the case of public safety communications: An integrated voice and data mobile network could provide first responders with a richer set of data, accelerating incident mitigation and improving emergency workers' welfare. However, there are issues that need to be settled before LTE Direct is ready for widespread use, including security and privacy concerns, revenue models, and frequency coordination. For these reasons, LTE Direct technology will not reach the consumer market space for another two to three years.

[Photo credit: First Responder Network Authority]

Device-to-device communication enables discovery of and direct communications between geographically close devices on a network [1]. LTE Direct is a D2D proximal discovery service driven by mobile technology and chip manufacturing firm Qualcomm. The technology is based on the Proximity Services (ProSe) technology standards set forth in 3rd Generation Partnership Project (3GPP) Release 12. Further enhancements for ProSe applications in general, including mission critical push-to-talk (MCPTT) for public safety communications, will be issued in 3GPP Release 13. ProSe was initially geared for public safety use, and LTE Direct could benefit this use case. LTE Direct could be used as an alternative to older public safety communication technologies that are at risk of becoming obsolete as second-generation (2G) mobile begins to sunset. Additionally, by using LTE Direct over existing mobile network operator (MNO) networks, emergency services networks would remain as up-to-date as the MNO's network, providing a dependable upgrade path [2].

ProSe communication

3GPP specifies two modes for ProSe: Network-authorized direct communication and network-independent direct communication. In network-authorized direct communication, user equipment (UE) always requires network assistance in establishing a communication link. Users are required to connect to the LTE network for timing, user authentication, and resource allocation. Once this initial setup occurs, the communications would be D2D only. Network-independent direct communication connections, authorized only for public safety communications, do not require assistance from the network to establish a communication link. This mode allows public safety officials to use one-to-one and one-to-many broadcasting without network infrastructure during emergencies. The specifics on exactly how these services will be launched from the device are unclear, but could entail being launched by the user from the device (much like Wi-Fi) or by connecting to static proximity beacons [2].

ProSe also has two distinct components: discovery and communication. LTE's air interface can "discover" and identify other devices in a given range of the user. Using ProSe technology, devices are capable of continuously sensing their surrounding environment to search for expressions [3], also referred to as affinity monitors, and broadcast data over an approximate range of 500 meters (m). Public expressions are application-agnostic and can be decoded by any device. Private expressions are application-specific and can only be decoded by devices with a key [4].

Although the ProSe communications feature is more useful for public safety users, discovery services do carry commercial potential. By implementing ProSe features through LTE Direct, MNOs are able to offer a variety of ambient awareness services to both the consumers and retail companies. Discovery also allows users to set expressions to broadcast details such as sales and services or search for nearby activities of interest [4, 5].

[Photo credit: First Responder Network Authority]

Public safety networks

Many countries are exploring using LTE for public safety, and LTE Direct could be a part of these emergency services networks. In the US, a 20 megahertz (MHz) band of 700 MHz spectrum has been allocated for public safety services and is dubbed the First Responder Network Authority (FirstNet). The initiative addresses a key recommendation of the 9/11 Commission regarding communications used by police, firefighters, and emergency medical personnel. In March 2017, AT&T was selected to build, operate, and maintain the nationwide wireless broadband network that will support FirstNet [6]. As 5G capabilities advance in the coming years, the two organizations will collaborate to ensure that the network is equipped to deliver data and video to first responders at exponentially increased speeds [7]. FirstNet and AT&T's first order of business is building the core network and delivering individualized state plans to US states and territories detailing the proposed network development in their jurisdictions; finalized plans are slated to be delivered in late 2017 [8]. Projections indicate that FirstNet will be "substantially in operation" by 2022 [9].

The Asia-Pacific region is also expected to have a large influence on the adoption of public safety LTE communications. The Korean government has shown strong support for the establishment of a national public safety network operating on a dedicated spectrum [10]. In June 2016, South Korean telecommunications operator KT Corp announced a trial public safety LTE network in preparation for the 2018 Winter Olympics. The spectrum that will be used for this network is 718 to 728 MHz uplink and 773 to 783 MHz downlink. KT built wireless base stations and provided special handsets designed for communication in the network [11]. A 2015 ABI report stated that China was actively testing LTE-TDD (time division duplex) for public safety communication and had allocated 20 MHz in the 1400 MHz band. Several pilot tests were carried out in Beijing, Tianjin, Nanjing, and Shanghai. China was also promoting a new protocol in partnership with Huawei and ZTE, as well as forming an LTE-based broadband trunking communication alliance called B-TrunC. LTE public safety networks were also planned and/or trialed in other countries, including Germany, Canada, the UK, Belgium, Qatar, the UAE, and Australia [12].

LTE Direct advantages and adoption

Key attributes of LTE Direct include:

- Will work on licensed LTE spectrum (700 MHz, 1700-2100 MHz, 1900 MHz, and 2500-2700 MHz).
- Devices can communicate without network assistance (bypassing evolved base stations, or eNodeB) or with network assistance (connecting to eNodeB) [13].
- Up to 300 megabits per second (mbps) downlink and 75 mbps uplink data rates [13].
- MNOs can use TDD or frequency division duplex (FDD) for UEs to broadcast or listen for expressions within device range.
- Reduce network congestion by offloading cell traffic [5].

Although it is unlikely LTE Direct will enter service within the next two to three years, the following technical advantages will drive adoption of LTE-based D2D technologies [2, 4]:

- Power efficiency: Existing ProSe technologies are not yet as power efficient as LTE Direct is anticipated to be. LTE Direct handles discovery at the device level without battery-draining network pings.
- One-to-one and one-to-many functionalities: LTE Direct ProSe enables public safety agencies to leverage one-to-one and one-to-many D2D functionalities and also provides the capability

TABLE 1. LTE Direct Competing Technologies

Standard	LTE Direct	Wi-Fi Direct	Wi-Fi Aware	Bluetooth Low Energy
Range	500 m	100 m–200 m	100 m–200 m	50 m
Spectrum	LTE licensed spectrum	2.4 gigahertz (GHz) and 5 GHz unlicensed spectrum	2.4 GHz and 5 GHz unlicensed spectrum	2.4 GHz and 5 GHz unlicensed spectrum
Data Rate	1 gigabit per second (Gbps)	250 mbps	1 Gbps	24 mbps
Pros	• Part of global standard • Built with application interoperability in mind • Low power consumption	• Only one connecting device needs to be Wi-Fi Direct-compatible to establish connection • Device manufacturer does not affect connectivity • Does not require a wireless access point	• Software able to run over Bluetooth beacons • Low power consumption	• Localized area • Privacy
Cons	• Challenges in establishing cross-network communication	• Uses Wi-Fi Protect Setup, which is vulnerable to brute force attacks	• Localized indoor area	• Proprietary • Power consumption

for public safety networks to leverage commercial LTE network growth.

- Network efficiency: Capability for MNOs to enhance capacity, coverage, and efficiency in mobile networks by enabling D2D at the edge of cells, potentially increasing throughput to 65% and creating space for better reuse of spectrum.
- New revenue sources for MNOs: LTE Direct ProSe adds a new variety of revenue prospects for MNOs like subscription fees for enabling D2D services, application program interface (API) access fees for application service providers (ASP), and fees generated by providing value-added services for retailers (advertising).
- Privacy: Devices are not required to reveal their location or allow location tracking to search for expressions, although discovery and communications features may require user-specific information to deliver tailored services.

LTE Direct challenges and concerns

Although heavily touted by Qualcomm and others as a potential commercial coup for operators, LTE Direct faces several challenges on its path to market adoption. More broadly, MNOs have yet to determine how they will charge for LTE Direct services—whether per user, per discovery event, per D2D connection, or by the amount of data used. Network management is also a sticking point as MNOs will need to come together and establish rules for frequency use and acceptable data rates when users are discovering "intra network." There is also the matter of dealing with potential interference between LTE Direct and LTE when users are in the same band. Last, and certainly not least, privacy and security concerns need to be mitigated. For instance, when using LTE Direct for discovery, certain information about users and their devices must be shared and that information must be protected. Concerns from a security perspective

[Photo credit: First Responder Network Authority]

include identity theft, spoofing, and denial of service attacks launched by receiving devices that send bulk download requests to sending devices [2].

The technology also faces competition from other established ProSe technologies, for example Wi-Fi Direct and Bluetooth Low Energy. The following table provides a comparison of LTE Direct and its competitors:

Outlook

In August 2015, Gartner pegged LTE Direct's market penetration to be less than 1% of the targeted audience and concluded that the technology was still in its infancy. However, companies backing LTE Direct have plans to increase the market share by using the technology for purposes beyond pushing alerts to consumers. For example, the technology could lay the foundation for vehicle-to-vehicle (V2V) communications. D2D communications like LTE Direct are expected to play a large role in the new 5G mobile standards by enabling wireless communication between a host of devices with limited reliance on mobile networks. The appeal in using LTE Direct for 5G D2D communications is its ability to transmit over long distances (500 m) at high data rates (1 Gbps). Additionally, LTE Direct-enabled chipsets will ship in 2017—in line with future 3GPP releases addressing the development of 5G technologies. LTE Direct also has the capacity to scale as 5G grows since the platform builds off existing LTE network infrastructure [2, 14, 15].

References

[1] Shen X. "Device-to-device communication in 5G cellular networks." *IEEE Network.* 2015; 29(2): 2-3. doi: 10.1109/MNET.2015.7064895.

[2] Bell S. "Device-to-device communications." Heavy Reading. September 2015.

[3] Expressions are 128-bit packages of data that broadcast activities of interest or services offered by the user.

[4] Qualcomm. "Creating a digital 6th sense with LTE Direct" [Presentation]. September 2015. Available at: https://www.qualcomm.com/media/documents/files/creating-a-digital-6th-sense-with-lte-direct.pdf.

[5] Lin X, Andrews J, Ghosh A, Ratasuk R. "An overview of 3GPP device-to-device proximity services." *IEEE Communications Magazine.* 2014; 52(4): 40-48. doi: 10.1109/MCOM.2014.6807945.

[6] FirstNet. "FirstNet partners with AT&T to build wireless broadband network for America's first responders" [Press release]. 30 March 2017. Available at: https://www.firstnet.gov/news/firstnet-partners-att-build-wireless-broadband-network-americas-first-responders.

[7] AT&T. "AT&T Selected by FirstNet to build and manage America's first nationwide public safety broadband network dedicated to first responders" [Press release]. 30 March 2017. Available at: http://about.att.com/sotry/firstnet_selects_att_to_build_network_supporting_first_responders.html.

[8] FirstNet. "Top 10 frequently asked questions." March 2017. Available at: https://www.firstnet.gov/mediakit.

[9] Kridel T. "LTE for public safety: Don't count on it." Heavy Reading 4G LTE Insider. April 2015; 6(2).

[10] Rehbehn K. Asia-Pacific illuminates the path toward public safety LTE networks. 451 Research. 27 October 2016. Available at: https://451research.com/report-short?entityId=90621.

[11] TeleGeography. "KT completes trial operation of PS-LTE infrastructure" [Press release]. 15 Jun 2016. Available at: https://www.telegeography.com/products/commsupdate/articles/2016/06/15/kt-completes-trial-operation-of-ps-lte-infrastructure/.

[12] Lian, JS. "Emergency response public LTE access." ABI Research. 23 July 2015.

[13] Adibi S. "A mobile health network disaster management system." In: *2015 Seventh International Conference on Ubiquitous and Future Networks (ICUFN).* 7-10 July 2015; Sapporo, Japan. doi: 10.1109/ICUFN.2015.7182579.

[14] 3GPP. "LTE-Advanced Pro ready to go" [Press release]. 28 October 2015. Available: http://www.3gpp.org/news-events/3gpp-news/1745-lte-advanced_pro.

[15] Fabre S. "Hype cycle for communications service provider infrastructure, 2015." Gartner Research. 4 August 2015.

FROM LAB TO MARKET

News from the NSA Technology Transfer Program

KAPALYA

Improving encryption via an NSA PLA

Honolulu company Kapalya, Inc. exclusively licensed NSA's patented authenticated encryption technology to position the start-up ahead of the competition. The company's system focuses on end-to-end encryption service using encrypted keys provided in real time. Using NSA patented technology, Kapalya will offer a solution that uses authenticated encryption for unstructured data on endpoints, corporate servers, and cloud servers, as well as a way to move these files securely between endpoints and servers while masking privileged users from viewing this data. Kapalya's new technology is cloud and carrier agnostic, allowing customers the freedom to use the device or platform of their choice.

The NSA Technology Transfer Program (TTP) began working with Kapalya in 2015, formally signing a patent license agreement (PLA) in March 2017. As a result of securing the exclusive PLA with NSA, Kapalya CEO Sudesh Kumar won a local shark tank event judged by five leading venture capitalists. Mr. Kumar is currently preparing to take his new technology to market.

> ***A game changer … that is how I describe the effect of licensing NSA technology on my business.***

When asked how the NSA PLA will affect his business, Mr. Kumar responded: "A game changer … that is how I describe the effect of licensing NSA technology on my business. The next generation of our data encryption app will have NSA's patented technology as the foundation. Once released commercially, this app will be cryptographically stronger and more efficient than existing authenticated encryption solutions in the marketplace."

The NSA TTP establishes partnerships between NSA and industry, academia, and other government agencies to help advance mission, foster innovation, and promote technology commercialization. For more information, visit www.nsa.gov/techtransfer.com.

[Photo credit: loops7/iStock/Thinkstock]

National Security Agency Central Security Service

Defending Our Nation. Securing The Future