An Investigation into Trust and Reputation Frameworks for Autonomous Underwater Vehicles

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy by

Andrew Bolster

June 2016

Contents

List of Figures

List of Tables

Preface

This thesis is primarily my own work. The sources of other materials are identifed.

Abstract

As Autonomous Underwater Vehicles (AUVs) become more technically capable and economically feasible, they are being increasingly used in a great many areas of defence, commercial and environmental applications. Increasingly, these applications are tending towards using independent, autonomous, ad-hoc, collaborative behaviour of teams or fleets of these Autonomous Underwater Vehicle (AUV) platforms. This convergence of research experiences in the Underwater Acoustic Network (UAN) and Mobile Ad-hoc Network (MANET) fields, along with the increasing Level of Automation (LOA) of such platforms, creates unique challenges to secure the operation and communication of these networks.

The question of security and reliability of operation has usually been resolved by having a centralised coordinating agent to manage shared secrets and monitor for misbehaviour. However, in the sparse, noisy and constrained communications environment of Underwater Acoustic Networks (UANs), the communications overheads and single-point-of-failure risk of this model is challenged (particularly when faced with capable attackers).

As such, more lightweight, distributed, experience based[1] systems of "Trust" have been proposed to dynamically model and evaluate the "trustworthiness" of nodes within a MANET across the network to prevent or isolate the impact of malicious, selfish, or faulty misbehaviour. Previously, these models have monitored actions purely within the communications domain. Moreover, the vast majority rely on only one type of observation (metric) to evaluate trust; successful packet forwarding. In these cases, motivated actors may use this limited scope of observation to either perform unfairly without repercussions in other domains/metrics, or to make another, fair, node appear to be operating unfairly.

This thesis is primarily concerned with the use of terrestrial-MANET trust frameworks to the UAN space. Considering the massive theoretical and practical difference in the communications environment, these frameworks must be reassessed for suitability to the marine realm. We find that current single-metric Trust Management Frameworks (TMFs) do not perform well in a best-case scaling of the marine network, due to sparse and noisy observation metrics, and while basic multi-metric communications-only frameworks perform better than their single-metric forms, this performance is still not at a reliable level. We propose, demonstrate (through simulation) and integrate the use of physical observational metrics for trust assessment, in tandem with metrics from the communications realm, to improve the safety, security, reliability and integrity of autonomous UANs.

[1] rather than "Evidence based" in the case of shared keys, Public Key Infrastructure (PKI) etc.

Three main novelties are explored and demonstrated in this work; Trust evaluation using metrics from the physical domain (movement/distribution/etc.), Demonstration of the failings of Communications-based Trust evaluation in the sparse, noisy, delayful and non-linear UAN environment, and the opportunity of Multi-Domain Trust assessment across physical and communications domains.

The final two elements of this novelty apply a machine learning methodology to establish selection filters for abstract and cross-domain / metric misbehaviours. This is demonstrated through a series of simulated experiments modelling the underwater communications and kinematic environments and we hope that these hypotheses, results, and novelties can be tested and proven in practical experimentation in the near future.

FIX: come back to the abstract

Acknowledgements

There are many people who deserve the highest thanks for their support, patience, kindness and understanding. The greatest thanks have to be distributed among my family and friends, for putting up with my madness; including but not limited to the madness of starting it, the madness of following the project to a different city and institution midway through, and the madness of seeing it through. Maybe I'll get a job that can actually be explained! I beg the indulgence of my examiners, supervisors, advisers and colleagues, as this has been a fascinating collection of areas to explore, and while I'm aware that by looking at all of them I've done none of them justice, but it's been a hell of a ride.

Finally, I must thank Professor Marshall, without whom this work wouldn't have been attempted let alone completed. Any mistakes are his and he can be contacted at alan.marshall@liverpool.ac.uk for comments and corrections.

Alan-hu Akbar

Chapter 1

Introduction

1.1 Mobile Ad-hoc Networks (MANETs)

With the explosive growth in the use of mobile telephony and the increasing miniaturisation and efficiency gains of portable communications devices, the classical paradigm of broadcast/receiver (or server/client) communications has given way to an increasing use of decentralised, ad-hoc networks that take advantage of this network dynamism to improve service efficiency.

Whether these networks are decentralised cellular / Radio Frequency (RF) / 802.11 WiFi networks for use in disaster relief areas [**?**] or biologically inspired wireless sensor networks for low-energy, low-maintenance environmental monitoring [**?** **?**], MANET theory developed over the past 30 years has gone from its first formal definition, emerging from Defence Advanced Research Projects Agencys (DARPAs) Packet Radio Network research, to being an integral part of modern practical communications[**?**].

Minimally, a MANET consists of of a collection of mobile physical entities (nodes) that communicate cooperatively to collect, distribute, disseminate, and collate data and/or influence across an area. In most cases MANET nodes incorporate bi-directional transceivers to send and receive data[1] MANETs may utilise omnidirectional, static, or steerable communications antennae, and a selection of protocols such as WiFi, Bluetooth, Global System for Mobile communications (GSM), Universal Mobile Telecommunications System (UMTS), as well as Optical or Acoustic media, and may incorporate a range of mobilities across nodes, from static devices, terrestrial and marine surface platforms, as well as aerial and underwater platforms. A core characteristic of MANETs is the inclusion and integration of heterogeneous node collections, i.e. different nodes or groups of nodes in a network may have different capabilities in terms of propulsion, sensor apparatus, communications capability, etc.

MANETs may be totally independent with no external connections, may include independent per-node communications backhauls (e.g. Cellular Modems in mobile phones as part of a Bluetooth Personal Area Network), or include static nodes that provide infrastructure based backhaul. However, this multiplicity of variations and options presents several challenges to users and operators; the physical topology of MANETs can vary wildly over short periods of time.

[1] However this bi-directionality is not always a requirement; for example in the area of Wireless Sensor Network (WSN) [**?**])

A particular challenge to MANET operation is that given any node may operate as a routing / gateway node, if/when that node moves to a different region, network segments that had previously used that node as a routing path must renegotiate / re-establish their routes. These situations, if not appropriately managed, lead to opportunities for subversion and selfishness.

The characteristics of MANETs as defined by Corson et al. are paraphrased in Table **??**.

Table 1.1: Summary of Characteristics of MANETs[**?**]

Dynamic Topologies	Nodes are free to move arbitrarily; thus, the typically multi-hop network topology may change randomly and rapidly at unpredictable times, and may consist of both bidirectional and unidirectional links.
Bandwidth Constrained, Varied Capacity	Wireless links will continue to have significantly lower capacity than their hardwired counterparts. In addition, the realized throughput of wireless communications, after accounting for the effects of multiple access, fading, noise, and interference conditions, etc., is often much less than a radio's maximum transmission rate. One effect of the relatively low to moderate link capacities is that congestion is typically the norm rather than the exception, i.e. aggregate application demand will likely approach or exceed network capacity frequently.
Energy Constrained Operation	Some or all of the nodes in a MANET may rely on batteries or other exhaustible means for their energy. For these nodes, the most important system design criteria for optimization may be energy conservation.
Limited physical security	Mobile wireless networks are generally more prone to physical security threats than are fixed cable nets. The increased possibility of eavesdropping, spoofing, and denial-of-service attacks should be carefully considered. Existing link security techniques are often applied within wireless networks to reduce security threats. As a benefit, the decentralized nature of network control in MANETs provides additional robustness against the single points of failure of more centralized approaches.

1.1.1 MANETs in Harsh Environments

As MANETs grow beyond the terrestrial arena, their operation and the protocols designed around them must be reviewed to assess their suitability to different communications environments, ensuring their continued security, reliability, and performance.

The distributed and dynamic nature of MANETs mean that it is difficult to maintain an evidence based "trust" system such as Trusted Third Party (TTP), Certificate Authority (CA) or using PKI. In both cases, there is the assumption of a run-time canonical source of trust, i.e. a "Master" node or Certifying Authority that can objectively coordinate the security and trust of the network. This single-point-of-failure is antithetical to MANET architectures, and given the normally limited transmission, storage, battery and computational power of MANET nodes, the overheads of true TTP or PKI architectures have been out of the realms of practicality for most applications. Therefore, a distributed, collaborative system must be applied to these networks[2].

[2] **?**] have demonstrated an intriguing low-power Distributed CA based MANET architecture, however given the soon-to-be-discussed assumptions about capable attackers(**??**), this "semi-decentralised" approach is less than ideal

Such distributed TMFs aim to detect, identify, and mitigate the impacts of malicious actors by distributing per-node assessments and opinions to collectively self-police behaviour. As such, TMFs can be used to predict and reason on the future interactions between entities in a system.

TMFs provide information to assist the estimation of future states and actions of nodes within MANETs. This information is used to optimize the performance of a network against malicious, selfish, or defective misbehaviour by one or more nodes. Previous research has established the advantages of implementing TMFs in 802.11 based MANETs, particularly in terms of preventing selfish operation in collaborative systems, and maintaining throughput in the presence of malicious actors [**?** **?**].

These works have focused on operations in the communications domain, usually relying on one type of observation or metric; Packet Loss Rate (PLR) or successful forwarding of nodes. Given the increasingly multi-factor nature of MANET security and integrity concerns, these style of frameworks do not look at information outside of their domain or even their metric. This exposes significant parts of the overall systemic threat surface to unobservable vulnerability, reducing the ability for a system to be "trusted" (**??**).

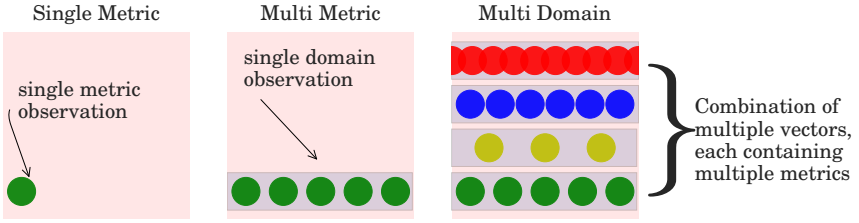


Figure 1.1: Multi-Domain Threat Surface

1.2 Autonomous Systems Approach to Trust and Trust Engineering

1.2.1 Autonomous MANETs

Autonomy is the capability of an entity to assess its environment and make informed, un-coerced decisions. In the MANET context, this sliding scale of capability ranges from basic automated collision avoidance systems while under direct operator control[3], through to self-regulating mission guidance and execution, with limited human interaction[4]. This non-deterministic operation presents significant challenges to security and integrity. Fundamentally, previously accepted formal verification methodologies for guaranteeing operation are not currently capable of accurately validating the actions and interactions of a fleet or swarm of autonomous nodes in

[3]For example Automated Driver Assistance Systems entering the consumer vehicle market through the likes of the Tesla P85 and Ford Kuga [**?**]

[4]No such systems have been actively deployed, but this form of collaborative autonomy is the centre of much commercial and academic research[**?** **?** **?**]

dynamic, noisy environments with imperfect operators [**?**]. As such an overlapping combination of "Secure" and "Trusted" approaches is required throughout the lifecycle and operation of an Autonomous MANET capability to maintain the integrity of such collaborative systems.

1.2.2 Trust vs Security vs Integrity

Early attempts to secure and protect the integrity of MANETs have relied on various forms of strong-cryptography to protect information being transferred from tampering or malicious inspection. While such approaches protect the integrity of individual pieces of data, the increased computation, and storage requirements of modern, strong, decentralised cryptographic systems presents a clear avenue for Denial of Service (DoS) attacks on MANETs. This threat is particularly relevant in resource-constrained networks, where one or more aspects of the environment are limited, be it available power, mobility, data storage, onboard processing, bandwidth, and channel resources such as capacity and delay. In such networks, where there is a requirement for security and/or integrity monitoring, strong-cryptographic methods present an entirely new opportunity to potential attackers.

One solution to the trade-off between DoS-protection, and security is the assessment of "trustworthiness" of nodes within a local network. "Trust" in this case is an assessment of capability of a node based on previously observed behaviour. Using this Trust to make simple routing decisions is significantly simpler and faster than strong-cryptographic methods, particularly in multi-hop networks or resource constrained networks [**?**]. With Trust being reliant on the runtime awareness of some behaviour, and cryptography on the pre-establishment of some entropy store and the repeated reinforcement of that numerical security, they represent two very different approaches to system integrity with very different costs/benefits and in practice, some elements of both methodologies will be used in different contexts and applications as those applications dictate.

1.2.3 Systemic Trust and Trusted Development

As will be discussed further in **??**, the "Trust" in a system is critical well before a system is activated; the incubation, specification, design, development, production and testing of a system (particularly a system with some LOA or other non-deterministic operation) is critical to the Trust that an end user can put into that system, and particularly, how much Trust can be exhibited within and between that systems individual components.

1.2.4 Trust Operation Against Capable Attackers

In any security situation, the hazards and risks of a systemic vulnerability being identified and exploited by an attacker are tightly coupled to the expectation of capability of that attacker. Within the defence context of this work, it is assumed that any attacking agency has the complexity and resources of a nation state.

This is also one of the primary motivators for this particular direction of work; where increasingly complex and subtle evidentiary security measures (passwords, encryption, etc) are applied, with increasing pressures in computation, connectivity, or communication, the

assumption that a slightly-higher-technical-investment will protect a system from state-level espionage or infiltration (or the discovery of some technical flaw in the system) is unfounded, with many examples of cryptographic applications being "disseminated" through human or technical failings, both internal and external.

1.3 Maritime Autonomy

Given the physical difficulties and requirements of having humans operate underwater (particularly at depth), and the operational limitations of tethered Remotely Operated Vehicles (ROVs), or even untethered remote controlled surface platforms, there has been a great deal of research and commercial interest in the development of autonomous systems, particularly in the defence and petrochemical sectors, where the lives of human operators are most at risk in normal activity[**?**].

With potential applications ranging from the replacement of human divers in Mine-Counter Measure (MCM) activity, to persistent littoral hydrography for environmental flood plain monitoring, many aspects of the proposed efficiencies gained from moving to autonomous agents rely on decentralised, MANET style collaboration between individual agents, and a level of trust that an operator (or indeed, agents within the system itself) can have in terms of the current activity, readiness, and performance of such autonomous systems[**?**].

This "Trust", that we know implicitly in the human space, is particularly difficult to establish and maintain in the underwater environment, and as such, the development of stable, reliable, adaptable trust systems that can operate in this and other challenging environments is a limiting factor on the adoption of generalised distributed autonomous systems in the defence space[**?**].

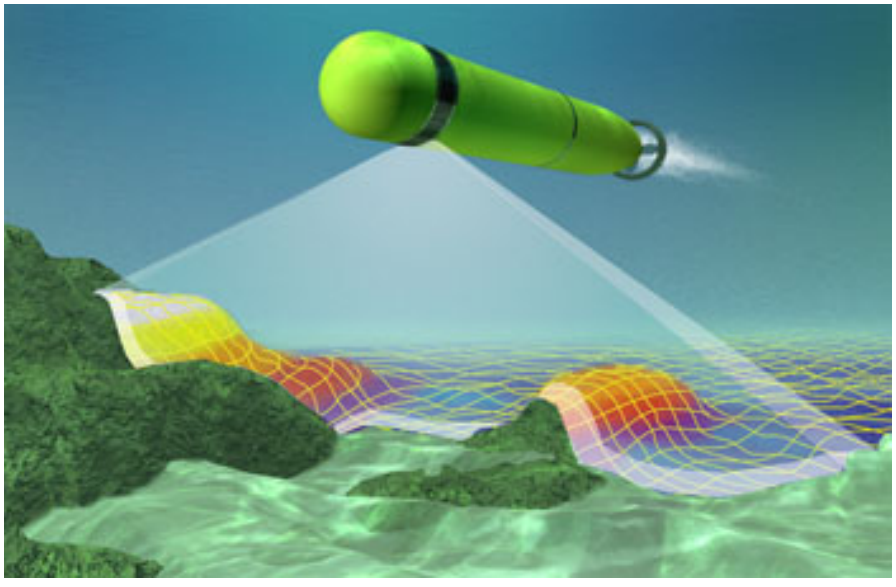


Figure 1.2: The AUV maps the seafloor with sonar©MBARI 2009

1.4 Thesis Layout and Contributions

1.4.1 ?? ??

In this chapter the current literature and research on the concepts, theory, and applications concerning Trust and Trust Management are explored, specifically leaning towards the applications of Trust within Autonomous MANETs.

In **??**, the abstract quantity of "trust" is explored, In **??**, Autonomy and "Trusted Operation" of autonomous systems is investigated from a system architects and a system operators perspective. In **??**, current use and applications of Trusted operation of MANETs is explored, including current TMFs.

1.4.2 ?? ??

In **??**, the maritime context is investigated, particularly the mechanisms of maritime acoustic communications, as well as the opportunities and challenges of the marine acoustic channel and its modelling (**??**). Additionally, the application scope of AUVs, littoral and sub-marine operations are explored to provide context to the problem (**??**).

1.4.3 ?? ??

In **??**, the need for multi-metric trust assessment in UAN is demonstrated as an example of a harsh network environment.

The operation of a selection of traditional MANET TMFs is investigated in this environment. These challenges are characterised and results are presented that demonstrate a multi-metric approach to Trust can greatly enhance the effectiveness of TMFs in these environments.

In **??** an experimental configuration for the marine space is established, and the scenarios and results presented in **?**] are reviewed for comparison. In **??** findings in trust establishment and malicious behaviour detection are presented, comparing with current single metric TMFs (Hermes and Objective Trust Management Framework (OTMF)) and the use of this multi-metric (vector) approach to detecting malicious and selfish behaviour in autonomous marine networks is analysed using Multi-parameter Trust Framework for MANETs (MTFM).

The contributions of this chapter are the first study on the comparative operation and performance of TMFs in marine acoustic networks, and a discussion of metric suitability for TMFs in marine environments, informing future metric selection for experimenters and theorists, and identifying both the opportunity and need to generate trust from additional domains, such as the physical domain. Finally, methodology to assess the usefulness of metrics in discriminating against misbehaviours in such constrained, delay-tolerant networks is demonstrated.

Key parts of this chapter were published as .

?? ??

Current approaches to operational security have been focused on the establishment of trust/security in the communications domain, and ignore other potential threats to the network exploited

through physical movement. This threat is particularly evident in collaborative autonomous systems where nodes are tasked to accomplish some survey / exploration / observation objective in a distributed fashion, where individual nodes make decisions based on the actions of their "team". This collaboration opens the opportunity for a physically-misbehaving actor to selfishly conserve it's own resources, or maliciously "drain" a given target node. Current security / trust systems applied to MANETs are not concerned with the threat of such physical misbehaviours.

This chapter proposes a new approach to trust in resource-constrained networks of autonomous systems based on their physical behaviour, using the motion of nodes within a team to detect and potentially identify malicious or failing operation within a cohort. This is accomplished by looking specifically at operations within the three dimensions of the underwater space, based on kinematics of industry standard AUVs. A series of composite metrics based on physical movement are presented and applied to the detection and discrimination of sample physical misbehaviours. This approach opens the possibility of bringing information about both the physical and communications behaviours of autonomous MANETs together to strengthen and expand the application of Trust Management Frameworks in sparse and/or resource constrained environments.

?? ??

In this chapter a methodology is demonstrated that applies Grey Sequence operations and Grey Generators to provide trust assessment in a sparse, asynchronous metric space across multiple domains of trust. By utilising information from multiple domains, it is demonstrated that trust assessment can be more accurate and consistent in identifying misbehaviour than in single-domain assessment. Further, a methodology for assessing the usefulness of individual metrics in this cross-domain space is demonstrated, allowing for the elimination of redundant metrics, simplifying the runtime assessment process.

NORELEASE: IF CLASSIFICATION GOES SOUTH YOU NEED TO FIX THE INTRO

Key parts of this chapter are awaiting publication as .

Chapter 2

MANETs and Trust

2.1 Mobile Ad-hoc Network Topologies & Routing

MANETs are wireless networks consisting of mobile devices acting simultaneously as sensor/processing/effector nodes and routing nodes, acting without a classical Wireless Local Area Network (WLAN) structured network architecture. Given constraints on propagation in such wireless networks, it is impractical for nodes to be "fully connected" to the rest of the network, and is instead constructed from single-hop "node-pair" links, through and across which data is routed to more distant parts of the network. This link-wise approach coupled with inherant Node mobility results in a potentially highly dynamic topology, where the instantaneous graph of node-pair connectivity can change dramatically in short intervals, and it may take considerable time for the network to "re-optimise" for this new, possibly temporary, topology.

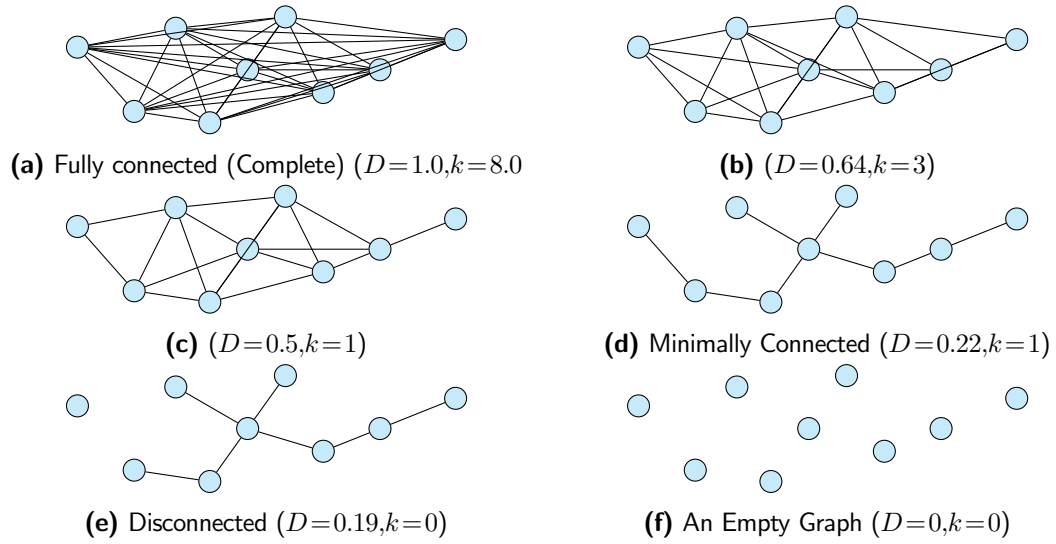
In order to understand and clarify the challenges to generic MANET applications, it is beneficial to explore some concepts from Graph Theory that apply to the discussion of MANET topologies.

2.1.1 Network Density and Connectivity

One fundamental compromise in the operation of wireless MANETs is the trade-off between the number of hops required between source and destination nodes and the effective bandwidth available to the network overall [**?**]. This compromise is encapsulated in the relative density of a given network; that is, the number of nodes in a given node's one-hop locality, drawing direct links between wireless transmission strength / reception sensitivity, the environmental

Graph Theory	Network Engineering
Vertex	Node
Edge	Link
Undirected	Symmetric
Directed	Asymmetric

Table 2.1: Basic mapping between Graph and Network Theory nomenclatures



(a) Fully connected (Complete) ($D=1.0, k=8.0$)

(b) ($D=0.64, k=3$)

(c) ($D=0.5, k=1$)

(d) Minimally Connected ($D=0.22, k=1$)

(e) Disconnected ($D=0.19, k=0$)

(f) An Empty Graph ($D=0, k=0$)

Figure 2.1: Network Density and Connectivity Examples

noise floor, environmental channel characteristics, the mobility of the nodes and the number of nodes deployed in a region.

From graph theory, the concept of "Density" is a $D_G = [0,1]$ bounded measure of how "Complete" or fully-connected a given graph G consisting of the set of vertices V and edges (links) E is. "Connectivity" can be generalised as the routing-ability and the possible redundancy of that routing-ability across the graph, i.e. for a "Connected" graph, such that there is a sequence of edges (a path) that can be traversed to link all possible pairs of nodes, but if any of the nodes were removed, the graph becomes disconnected or separated, that graph has a connectivity of 1. If this is not possible, but it is possible to disconnect the graph by removing two vertices, the graph has connectivity 2, and so on. A graph with a connectivity of 0 indicates that that graph has separated (or disconnected) vertices.

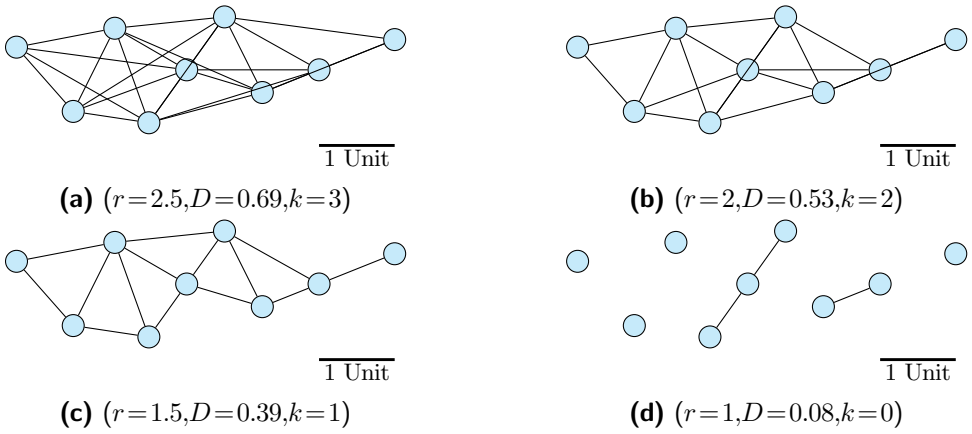
The concepts of graph density and connectivity are demonstrated visually in **??**, constructed with a uniform vertex distribution and decomposed by removing subsequent "longest edges" from a fully connected (Complete) graph.

2.1.2 Node Density in MANETs

In practical wireless network applications, network connectivity is a product of the physical distribution of nodes within an environment, and the propagation characteristics of the medium.

Taking the same physical configuration as in **??**, the structure, density and connectivity of the network with different assumed propagation ranges r can be shown (See **??**). Naturally, as the relative transmission range is reduced, the natural density and connectivity of the network reduces down to the point in **??** where the "network" is fundamentally broken.

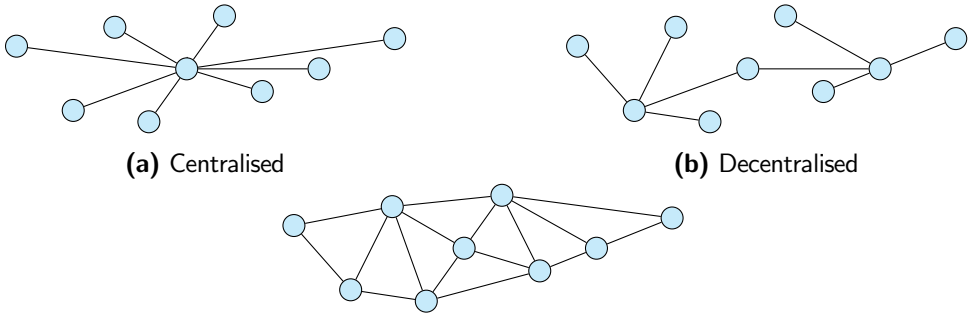
Another graph theoretic factor that is worth considering in MANET design is *Centrality*, that is, the measure of importance of individual nodes to the connectivity of the network. There are a great many methods for calculating centrality, however the important understanding from a network perspective is that of network-relative centralisation; in **??**, our previously deployed network



(a) $(r=2.5, D=0.69, k=3)$ **(b)** $(r=2, D=0.53, k=2)$

(c) $(r=1.5, D=0.39, k=1)$ **(d)** $(r=1, D=0.08, k=0)$

Figure 2.2: Examples of states of MANET topologies



(a) Centralised **(b)** Decentralised

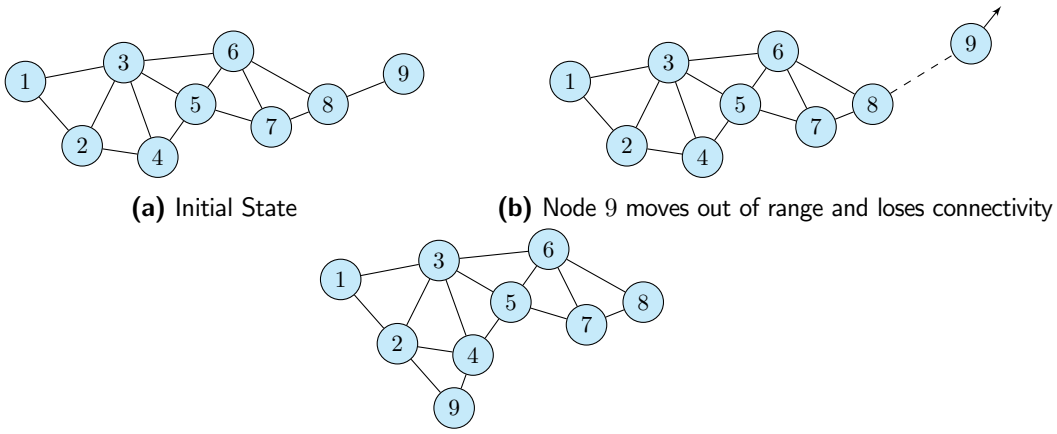
(c) Distributed

Figure 2.3: Example of routing strategies and logical connectivity in a sample network

layout is used to demonstrate three delineations of network centrality. In the first, all nodes are connected to one node, and that one node is totally responsible for communications and routing in the network. This is architecturally similar to switch-managed wired networks in a small office or home. In the second example, a generally de-centralised layout is set, with a generally hierarchical logical routing setup, where a few nodes take the brunt of the connectivity and bridging, almost creating centralised "sub graphs" with single uplinks. Finally, a truly "distributed" topology is shown, where there are no architecturally important nodes and, in general, shorter link distances.

The last issue to consider directly is that of the impacts of node mobility and temporary absence. In **??**, a node leaves the network from one "side", and later reappears at another part of the network. This is quite possible in cases where node links may be physically interrupted by obstructions (or indeed the planned paths for nodes are obstructed, requiring a course correction and delay). When the node reappears, all routing information that may have been established (i.e. node 5 would send packets destined to node 9 to nodes 6 or 7 rather than nodes 3 or 4) must be renegotiated and re-distributed across the network before efficient operation can continue. Secondarily to this, any authentication and validation that may be required to positively identify node 9 as *actually being* node 9 may have to be transited from node 8, which had previously been the only node directly communicating to node 9.

These factors of network density, link length, centralisation and mobility simultaneously give MANETs their strength and resilience as well as their risks and challenges.



(a) Initial State

(b) Node 9 moves out of range and loses connectivity

(c) Node 9 rejoins the network at some future point in a different topology

Figure 2.4: Example of node mobility in MANET topology changes

2.2 Routing in Mobile Ad-hoc Networks

Given the decentralised nature of MANET operations, routing protocols have been an active area of research since their inception [**?**]. This research is classified according to the strategies used for discovering, monitoring and updating routes within the network, and are usually grouped into three classes; proactive (or Table Driven), reactive (or On Demand) and hybrid protocols. A summary of the generalised characteristics of these classes is shown in **??**. Additional, these classes can be further rearranged, combined or augmented based on assumptions made about the structure of the baseline topology, i.e. flat, hierarchical or geographic) or by assumed constraints of the available resources of the nodes within the network i.e. heterogeneity in power, mobility, or communications capabilities where resource-based heuristics are used as well as purely topological considerations[**? ?**].

2.2.1 Proactive Routing

In Proactive routing, protocols attempt to maintain a up-to-date, global topology awareness of the network, where every node knows how the best next-hop to contact any other node in the network. This is extremely efficient for relatively small, static networks, with minimal storage and time requirements [**?**]. When the network topology is significantly modified by a shift in topology, either due to a node "dropping out" or moving, route renegotiation and optimisation is extremely resource consuming, as this global state is converged upon in a distributed manner by nodes exchanging their local knowledge of the "new" topology. The decomposition and updating of the node-knowledge of the network state, and the method of updating these state-tables, is the primary differentiator between proactive protocols, a selection of which are summarised below.

Destination-Sequences Distance Vector (DSDV)

Destination-Sequences Distance Vector is a loop free derivative of the Distributed Bellman-Ford algorithm where each node maintains two tables; one that attempts to maintain a globally accurate next-hop routing table for all destination nodes (the routing table) and a route

advertisement table, monitoring routes that the node itself can provide. These tables are updated both periodically and opportunistically. Loop-free status is maintained by monitoring a monotonic "sequence number", which guarantees that if a long-loop returned packet is observed, it is discarded in favour of a route with a higher sequence number (i.e. newer route) [**?**].

Optimised Link State Routing (OLSR)

Optimised Link State Routing reduces the traffic-overhead of truly distributed link-state exchange and monitoring by establishing a multipoint replaying strategy (MRP) where nodes select a subset of their one-hop network relay to retransmit their packets, based on the two-hop connectivity of the network, thereby reducing contention and overheads by reducing local re-transmitters. However, OLSR does not monitor link *quality* beyond binary "active/failed" state which can lead to non-optimial MRP and route selection in wireless networks for instance.

Topology Dissemination Based On Reverse-Path Forwarding (TBRPF)

Topology Dissemination Based On Reverse-Path Forwarding consists of two main modules; the neighbour discovery (TND) module and the routing module; the TND uses differential updates to report only the *changes* in the local topology. Further, instead of flooding the entire network with updates, TBRPF selectively updates the relevant route updates to nodes that are on the minimum spanning tree route of of that update. Full topology updates are also used on a periodic, but occasional basis to maintain consistency and visibility. Given the use of differential updating, TBRPF is more responsive and resilient in the face of dynamic mobile networks and incurs lower traffic overheads.[**?**]

2.2.2 Reactive Routing

In contract to Proactive Routing, Reactive (or "on-demand") routing establishes routing information when it is required, rather than in advance or periodically. This route establishment is usually based on a request-response exchange where the node requesting routing information "floods" its local network with next-hop requests. The structure of this flooding (and the context of any responses) are the main differentiators between protocols, discussed below. There are two main sub-classes of reactive routing; source routing and hop-by-hop routing where packet routing information is either totally planned in advance and encapsulated in the packet on transmission, or decided at each forwarding point respectively. The on-demand nature of of route discovery can lead to significantly lower traffic than proactive routing protocols, but this is often a trade-off between lower average traffic and larger pre-transmission discovery delays. As such, reactive routing lends itself to low-traffic, delay tolerant, dynamic mobile applications as it does not require rediscovery after every *topology* change, but only on transmission along a new or stale route.

Dynamic Source Routing (DSR)

On-demand route formation when a transmitting node requests one. However, packets include full routing information instead of relying on the routing tables at each intermediate device[**?**]. Effective for small to medium, minimally mobile networks due to inclusion of route caching which reduces the number of route request discovery phases and associated congestion. Ineffective in large networks due to full-route packet overheads as network scale increases, and less than ideal for mobile networks due to cache-miss delays.

Ad hoc On-demand Distance Vector (AODV)

Based on DSDV and DSR; uses beaconing and sequence numbering (DSDV) as well as shortest path route discovery from DSR, except that AODV does not use full-path source routing, instead relying on intermediate-routing based only on a destination. This optimisation reduces overhead and is more resilient to highly dynamic deployments at the cost of variable and potentially very long delays due to slow route construction or complete retransmissions due to link failure.[**?**]

Routing On-demand Acyclic Multipath (ROAM)

ROAM uses inter-nodal coordination along directed acyclic sub-graphs, which is derived from the routers distance to destination. This operation is referred to as a diffusing computation. The advantage of this protocol is that it eliminates the search-to-infinity problem present in some of the on-demand routing protocols by stopping multiple flood searches when the required destination is no longer reachable. Another advantage is that each router maintains entries (in a route table) for destinations, which flow data packets through them (i.e. the router is a node which completes/or connects a router to the destination). This reduces significant amount of storage space and bandwidth needed to maintain an up-to-date routing table. Another novelty of ROAM is that each time the distance of a router to a destination changes by more than a defined threshold, it broadcasts update messages to its neighbouring nodes, as described earlier. Although this has the benefit of increasing the network connectivity, in highly dynamic networks it may prevent nodes entering sleep mode to conserve power.

Associativity Based Routing (ABR)

ABR extends classical source-routing by including a stability ("associativity") heuristic of the long-term link state between mobile nodes, ensuring that the least-mobile nodes are preferentially used for routing. Further, this heuristic is applied outward from destination rather than from the source, selecting only the "best" route, reducing the likelihood of packet duplication in the

mid-network. However this "associativity" measure requires periodic beaconing forcing all nodes to remain active. Finally, in the case where the "best" route fails through an in-the-air topology change, there is no in-network path recovery mechanism, and link discovery must be restarted[**?**].

Location Aided Routing (LAR)

LAR incorporates location information (usually from Global Positioning System (GPS)), and generates a heuristic based on either the distance from the current node *towards* the destination location, or the distance from the current node *away from* the original source, minimising and maximising this distance respectively.These methods limit control overheads and usually accurately determine the shortest path. However, in highly mobile networks this behaviour appears increasingly flood-like (similar to DSR and AODV), and the general requirement for highly accurate and timely positional information restricts the application of this protocol

Cluster Based Routing Protocol (CBRP)

CBRP uses a hierarchical clustering topology where each cluster has a cluster-head which coordinates routing within that cluster. As only cluster-heads coordinate routing across clusters, transmission overheads are minimised compared to other route distribution methods. However, the negotiation and maintenance overheads and propagation delays associated with hierarchical clustering make the network susceptible to temporary routing loops as nodes may have inconsistent residual routing information during cluster re-negotiation

2.2.3 Geographic Random Forwarding (GeRaF)

GeRaF is a geographic, on-demand, opportunistically routed protocol that could be considered an edge case of the "reactive" definition, in that it relies on additional knowledge about the environment to make routing decisions. It's basic operation is that a source node broadcasts a packet to be sent, including it's own location and the estimated location of the intended recipient. Intermediate nodes then reactively assess their own optimality in relaying based on maximum distance advancement from source and sink positions, and contend for the channel to receive the packet and rebroadcast it on. The core innovation of GeRaF is this intermediate / receiver contention policy[**?** **?**]. One downside of this distance-only priority heuristic is that it can induce long, zig-zag paths that can be sensitive to node mobility [**?**].

2.2.4 Hybrid Routing

Hybrid routing protocols combine selected elements from both Proactive and Reactive routing in an attempt to minimise weaknesses in the respective classes. In many cases, this takes the form

of a tiered or bounded choice between proactive and reactive approaches based, where inside a given "set" of nodes (either physically proximate or or via a directory-based subgrouping), lower latency, deterministic, proactive approaches are used, and outside or between such sets, reactive approaches are applied. These in general exhibit "better" performance on average, particularly with increasing numbers and expanding distributions of nodes, but the variability in performance may be significant, particularly when sets require updating due to physical mobility of nodes, or where key gateway nodes are removed from the network somehow.

Zone Routing Protocol (ZRP)

A true-blend of Proactive and Reactive policies; ZRP draws "Routing Zones" around nodes based on hop-distance, within which routing is made proactively, providing immediate local routes, and reactive, on-demand routes outside this distance. This significantly reduces local overheads and delays by reducing the scope of potential routes as those nodes on the edge of the zone. The control of this boundary point is a significant challenge to optimise with respect to overall network scale.

Distributed Spanning Trees (DST)

Based on a combination of Hybrid Tree Flooding and Distributed Spanning Tree shuttling on tree based clusters where each cluster has a root node acting as a configuration leader. Routing updates are passed through direct neighbours and "up" the spanning trees under the root; this leads to a highly responsive and low over head routing policy in highly dynamic networks.[**?**]

Distributed Dynamic Routing (DDR)

A tree protocol similar to DST except without the need for a root node; trees are constructed and maintained by periodic neighbour beaconing, where each node becomes the potential root of its own tree within the "forest" of the wider network. The construction of this forest follows six phases; neighbour election, forest construction, intra-tree clustering, inter-tree clustering, zone naming and zone partitioning. Each of these phases are executed based on information received in the beacon messages. One of the strengths of DST is its lack of centralisation or a-priori structure requirement (i.e. root/clusterheads or static zone maps), however there is no equalisation method to balance the case where a gateway node that is a preferred neighbour to many subtrees becomes congested and represents a significant bottleneck, as the neighbour selection is predicated on graph connectivity alone, without taking maximum throughput into account. In variably connected networks this could potentially cause network-wide delays through mid-network packet drops.

Zone-based Hierarchical Link State (ZHLS)

Compared to ZRP, ZHLS extended the zoning concept to include some elements of LAR, by constructing hierarchical non-overlapping zones based on physical location as well as connectivity. This location management is purely decentralised, with no explicit "zone-heads", eliminating single-point-of-failure concerns and significantly reducing invalid-flooding overheads, as topology updates only carry towards zones where the information is relevant. Shifting topologies within zones are tolerated cleanly as the internal zone-map is flat rather than hierarchical, and does not require re-computation or re-location as long as the node stays within a given "zone". This static map also leads to a significant disadvantage in that this zone-map must be pre-set, so are inappropriate for fully-mobile applications or applications with dynamic geographic boundaries [**?** **?**].

Scalable Location Update Routing Protocol (SLURP)

With a similar hierarchical non-overlapping zone structure to ZHLS, SLURP does away with global routing through a deterministic mapping of node identifiers to "Home" regions, such that any node attempting to communicate with a node, can directly calculate from what "Home" zone that node originated.As and when nodes leave their "Home" region, they feedback to that region their current location. Subsequently when that node is a destination for a packet, the routing query is automatically directed to the home region which can direct the source node as to the direction of its destination, upon which the source can start sending data towards the destination using a most forward with fixed radius (MFR) geographical forwarding algorithm. Once the data reaches the zone where the destination currently resides, source-routing is used internally to complete the route. This strategy works well for relatively static networks with some mobile nodes, or where node mobility is "slow", such as WSN, however it still relies on pre-programmes static zone maps as per ZHLS [**?**]

Focused Beam Routing (FBR)

FBR is built upon the concepts of GeRaF, which uses straight line distance-from-receiver minimisation, extending with a cone-based prioritisation metric rather than absolute distance, mitigating the zig zag effect of GeRaF and implicitly supporting partially overlapping simultaneous alternate routing, improving system redundancy and (usually) eliminating the need for retransmission [**?**]. This is further augmented with an adaptive open-loop power control system which both aids in energy consumption and in preventing spurious blocking of the channel. FBR is particularly well suited to constrained energy mobility networks with high mobility and high retransmission costs [**?**].

Table 2.2: Comparison of Routing Strategy Classes (from ?])

Class / Area	Proactive	Reactive	Hybrid
Routing Structure	Both flat and hierarchical structures are available	Mostly flat except CBRP	Mostly hierarchical
Route Availability	Always available if nodes are reachable	Determined when needed	Depends on the location of the destination
Control Traffic Volume	Usually high, attempt at reduction is made. e.g. OLSR, TBRPF	Lower than Global routing and further improved using GPS. e.g. LAR	Mostly lower than proactive and reactive
Periodic Updating	Yes, some may be conditional e.g. STAR	Not required, however some nodes may require periodic beacons. e.g. ABRs	Usually used within each zone or between gateway nodes
Mobility Handling	Usually updates occur at fixed intervals. DREAM alters periodic updates based on mobility	ABR uses localised broadcast queries, ROAM uses threshold updates, AODV routing uses local route discovery	Usually more than one path may be available. Single point of failures are reduces by working as a group
Storage Requirements	High	Dependent on number of nodes kept or required; usually lower than proactive protocols	Usually depends on cluster or zone size; may become as large as proactive if clusters are big
Delay Level	Short routes are predetermined	Higher than proactive	Short for destinations in the same zone/cluster as source. Inter-zone may be as large as Reactive protocols
Scalability	Up to 100 nodes; OSPF and TBRPF may scale higher	Source routing protocols; up to a few hundred nodes. Point-to-point may scale higher. Depends on level of traffic and levels of multihopping	Designed for up to or more than 1000 nodes

2.3 Trust Definitions, Perspectives, and Relationships

For a term that is so common in every-day speech, "Trust"[1] is a challenging discussion area, particularly given the wealth of proposed definitions (Table **??**).

Beyond these dry, vague, and often "fuzzy" definitions, there is a significant ontological conflict between the subjective and objective perspectives of trust; is "trust" an attribute of the actor performing a given action, or of the observer of such an action? Or indeed is trust itself an action upon a relationship between actors? Is it qualitative or quantitative? These questions have challenged philosophers, psychologists and social scientists for decades.

In human trust relationships it is recognized that there can be several domains of trust for example organizational, sociological, interpersonal, psychological and neurological [**?**].

These domains of trust are, from a human perspective, quite natural and are formed during the earliest stages of linguistic integration. This leads to recognisable deviations in the experiential concept of "trust" across cultures with differing linguistic histories. This has led to a wealth of work in the social sciences (as well as management schools across the world) in to how to develop, understand, and repair trust across cultural boundaries [**?**].

As such it is important to explore the following areas of trust definitions, the characteristics of trust relationships and the impact of topology on the information available to assess trust within an abstract network before approaching the application of Trust towards Autonomous Systems and finally to MANETs.

Table 2.3: Definitions of Trust

Definition	Source
Assured reliance on the character, ability, strength, or truth of someone or something.	Merriam-Webster
Firm belief in the reliability, truth, or ability of someone or something	OED
The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a articular action important to the trustor, irrespective of the ability to monitor or control that other party	**?**]
An expectancy held by and individual or a group that the word, promise, verbal or written statement of another individual or group can be relied upon	**?**]

2.3.1 Modelling Trust Relationships

?] proposed a model of trust that encapsulates generalised factors of perceived trustworthiness of a *trustee* in interpersonal relationships (Table **??**), accommodating a subjective trustworthiness and risk-taking potentiality on the part of the *trustor*. This formulation of trust allowed a wider

[1] As a point of notation, in this work "Trust" and "trust" are used interchangeably to refer to the concept, action, or belief of a specified trusting relationship. Where Trust is capitalised outside of grammatical convention, it is to emphasise "trust as a concept" rather than a particular value or relationship

discussion of the characteristics of trust relationships, both between individuals and within networks or communities.

Table 2.4: Factors of Trust (from ?])

Factor	Definition
Ability	Collection of skills, competencies, capabilities and characteristics that enable a party to have influence or action within some specific domain
Benevolence	The extent to which a trustee is believed to want to do good to or by the trustor beyond a selfish profit motive
Integrity	Acceptance or adherence to a common set of principals of operation that the trustor finds acceptable

As shown in **??**, Mayer primarily focuses on the Trustor's perspective and processes with respect to a give trust-based relationship. Three primary factors of perceived trustworthiness; based on previous outcomes, are assessed and synthesised along with the Trustor's own interanalised propensity to Trust with respect to the different factors observed, to generate a given trust value. This trust value is incorporated with the risk / reward as assessed by the trustor to conclude what level of risk taking (Trust) can be assumed in the relationship between this trustor and a given trustee.

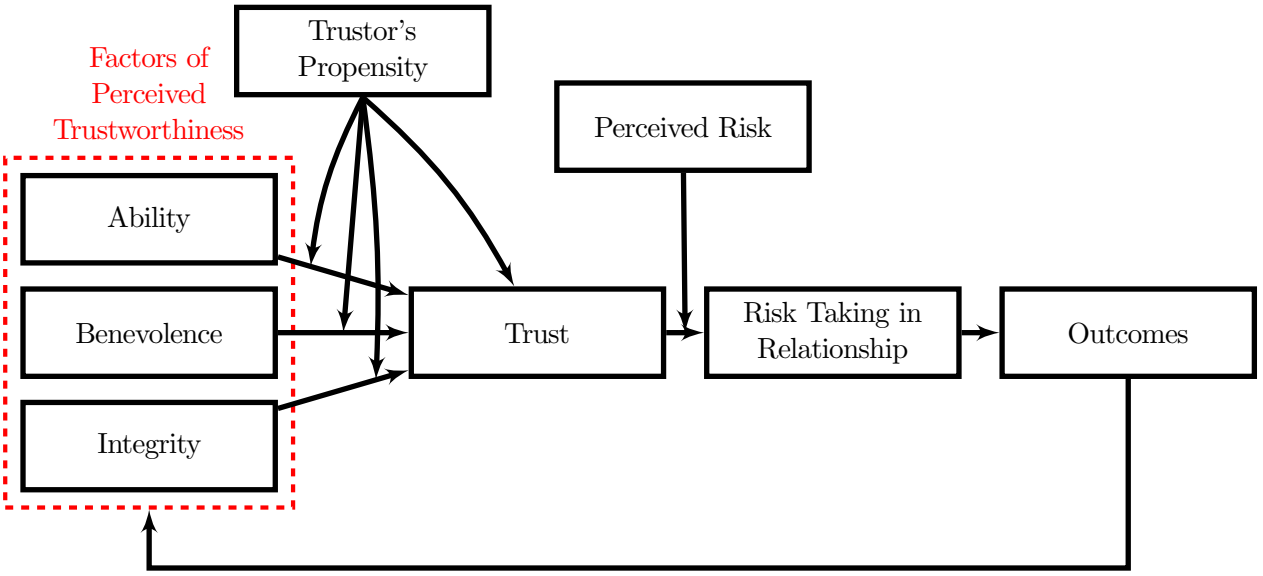


Figure 2.5: Model of Trust (from ?])

?] extended and synthesised Mayer et al's approach to personal and interpersonal trust towards a generalised concept of trust for human and autonomic/autonomous systems with alternative contextual definitions shown in **??** (including their approximate mappings to Mayer et al's approach.

?] suggests that there are two overarching forms of trust:

- Behavioural: That one entity voluntarily depends on another entity in a specific situation
- Intentional: That one entity would be willing to depend on another entity

Table 2.5: Factors of Trust for Autonomous Systems (from ?])

Factor	Definition	Mayer Term
Performance	The current and historical operation of the automation, including characteristics such as reliability, predictability, and ability	Ability
Purpose	The degree to which the automation is being used within the realm of the designers intent	Benevolence
Process	The degree to which the automation's algorithms are appropriate for the situation and able to achieve the operators goals.	Integrity

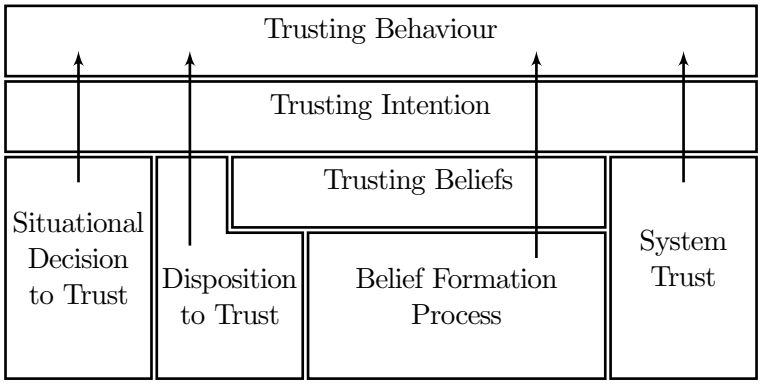


Figure 2.6: Trust Construct Relationships (from ?])

It is suggested that these overarching forms are supported by and indeed are drawn from four major constructs within social and networked environments, as identified by **?**]:

- Trusting Belief: the subjective belief within a system that the other trusted components are willing and able to act in each others' best interests
- Dispositional Trust: a general expectation of trustworthiness over time
- Situational Decision Trust: in-situ risk assessment where the benefits of trust outweigh the negative outcomes of trust
- System Trust: the assurance that formal impersonal or procedural structures are in place to ensure successful operation.

Sun argues that only System Trust and Behavioural Trust are relevant to trusted networking applications. However, it is arguable that in any communications network where the operation of that network is not the only concern, or where that network has to interact with any operator, then all of these factors come into play; as we will see (**??**). Both System and Behavioural trust rely on what Sun calls a "Belief Formation Process", or a trust assessment, while the other trust constructs deal with the interactions between trust and decision making against an internal assessment of network trustworthiness.

2.3.2 Taxonomy and Notations of Trust

To abstractly discuss Trust and trust modelling, **?**] present a $T\{\text{subject}:\text{agent},\text{action}\}$ notation for individual trust relationships, where *Subject* and *Agent* usually representing individuals

but may include groups of individuals or the network as a whole, and *Action* may be an action performed by a given agent or a property possessed by that agent.

This notation is normally abbreviated such that

$$T\{A:B,a\}=T^{a}_{AB} \tag{2.1}$$

where T^{a}_{AB} denotes the expectation or trust that a node A has that node B will successfully perform action a.

A special extension is assumed for multi-party trust relationships for the "action" of recommending another nodes perspective on a given nodes expectation or trust that it will perform an action a; (a_R)

$$T\{A:B,a_R\}=R^{a}_{AB} \tag{2.2}$$

Where the action under discussion is implied or there is a single action under debate (for example, packet routing in MANETs), the action superscript can be left out; T_{AB},R_{AB}.

Trust can be propagated through a number of nodes, forming a chain of assessments and recommendations, expressed as per (**??**). While the particular algorithms and methods of combination vary between TMFs, the $\cdot$ symbol is used as a generic operator. In any case, if a given node A has no knowledge about node B, or likewise B having no knowledge of C, trust between A,C is zero; $T_{ABC}=0$

$$T_{AC}\mapsto T_{ABC}=R_{AB}\cdot T_{BC} \text{ where } R_{AB}\neq 0,T_{BC}\neq 0 \tag{2.3}$$

Either notation; T_{AC},T_{ABC} is valid for this chaining depending on the context, where, T_{AC} is preferred where there is either one unambiguous transiting node B, or where the value is the multi-path trust synthesis across many individual links, where T_{ABC} is then the preferred notation for a single path within a multi-path network.

For discussion of individual links or subsets of links, Set notation can be applied to T_{AC} multi-path networks based on the graph theoretic notation introduced in **??**, for example;

$$T_{AC}\subseteq T_{AxC}\forall x|\overrightarrow{AxC}\in E \tag{2.4}$$

Similarly, these multi-path sets can be decomposed into their individual links, i.e. $T_x\geq 0\forall x\in T_{AC}$

Characteristics of Trust Relationships

There are five commonly considered characteristics or attributes of Trust relationships in general, but not all relationships exhibit them and they are not assumed to be a complete specification of Trust (synthesised from [**? ? ? ?**]):

- *Multi-Party* - One-to-one; one-to-many; many-to-one; many-to-many. Trust is not an absolute characteristic of a lone individual. Trust may include multi-agent abstractions (one-to-many), such as a preferential trust/distrust towards a group exhibiting a particular attribute, e.g.

members of the armed forces / police services. Likewise, there can be trustor/trustee attributes that can generalise relationships between collectives (many-to-many), e.g. Jets and Sharks[**?**].

- *Transitive* - Trust assessments can be shared (i.e. recommendations), where this second order trust assessment incorporates both the observed trustworthiness of the trustee, as well as the trustworthiness of the intermediate trustor. In some models this is further extended to include out-of-network intermediate trustors that have some other defined authority, e.g. PKI Certificate Authority
- *Evidential* - Trust must be based on some form of evidence-based observation or assessment, such as historical success rates of performing a certain action, or second-hand observations of trust from a third party.
- *Directional Asymmetry* - The majority of relationships are bi-directional but are asymmetric, i.e. between two entities who "trust" each other, there are two independent trust relationships that may have very different "values" or extents.
- *Contextual* - Trust can be variable and loosely coupled between contexts with respect to the action being assessed or the environment within which the trustee is operating, e.g. Doctors are trusted to perform medical procedures but that trust may not improve their success at correctly wiring an electrical plug. However there are plenty of counter-examples to this, as from [**?**], two of the three listed factors of trust are "Benevolence" and "Integrity" and these are unrelated to the ability of a trustee to perform a particular action, so it is reasonable to make an initial assumption that if a trustee is being benevolent in one activity or context, that that benevolence *should* extend to other contexts.
  [**?**] summarises these attributes in a series of axioms

Fundamental Axioms of Trust

?] demonstrate that by taking an Information Theoretic approach to trust as function of entropy and as such, uncertainty (i.e. trust having both a valency and a confidence) [2], a series of axioms can be constructed to model the interactions between trusting agents. Many of these axioms are mirrored in pure information theory, as discussed in [**?**].

*Axioms and their quoted descriptions from [**?**]*

Axiom 1: Concatenation propagation of Trust does not increase Trust

> When the subject establishes a trust relationship with the agent through the recommendation from a third party, the trust value between the subject and the agent should not be more than the trust value between the subject and the recommender as well as the trust value between the recommender and the agent.

This axiom sets up the rules for trust propagation from an entropic perspective and using the notation discussed above, can be expressed as in (**??**) (see **??** for context)

$$T_{AC} \leq \min(R_{AB}, T_{BC}) \tag{2.5}$$

[2] Donald Rumsfelds famous 2002 "Known Knowns" quote is a perfect example of this

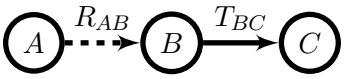


Figure 2.7: Trust Chaining

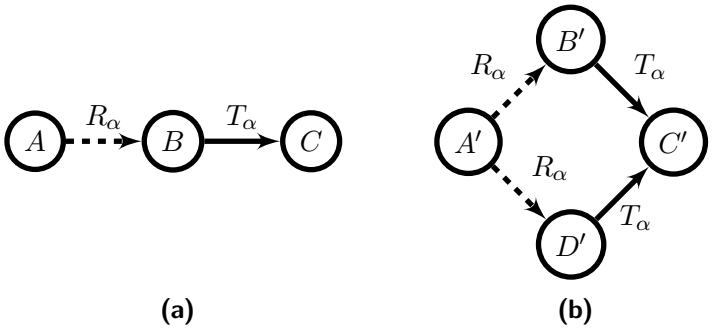


Figure 2.8: Trust Combination

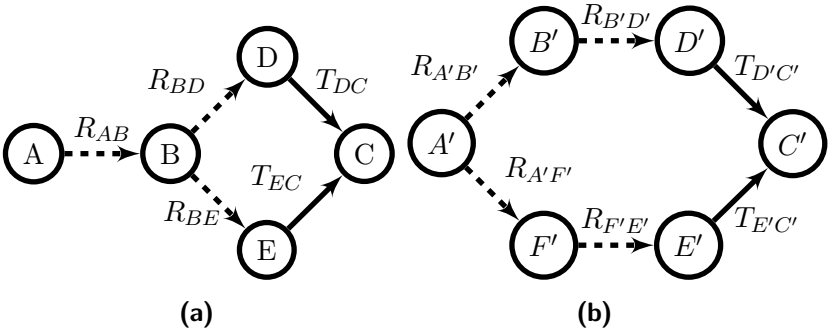


Figure 2.9: Trust Paths

Axiom 2: Multipath propagation of Trust does not reduce Trust

> If the subject receives *the same* recommendations for the agents from multiple sources, the trust value should be no less than that in the case when the subject receives fewer recommendations.

In this case, this axiom sets the groundwork for multi-node analysis of trust networks, shown in **??** and described in (**??**). In essence, adding additional information sources of the same observation should increase the trust value arrived at from a smaller subset of sources.

$$\begin{aligned} &T_{AC} \geq T_{A'C'} \geq 0, \text{ for } R_\alpha \geq 0, T_\alpha \geq 0 \\ &T_{AC} \leq T_{A'C'} \leq 0, \text{ for } R_\alpha \geq 0, T_\alpha \leq 0 \end{aligned} \tag{2.6}$$

Axiom 3: Trust based on multiple recommendations from a single source should not be higher than that derived from independent sources.

> When the trust relationship is established jointly through concatenation and multi-path trust propagation, [...] recommendations from independent sources can reduce uncertainty more effectively than can recommendations from correlated sources.

This axiom addresses the information independence of trust links; In **??** there are two networks, both with two potential chains from right to left-nodes (A,C,A',C'),.

$$T_{AC} = \{T_{ABCD}, T_{ABEC}\} \tag{2.7}$$

$$T_{A'C'} = \{T_{A'B'D'C'}, T_{A'F'E'C}\} \tag{2.8}$$

Given that the sub-chain T_{AB} prepends both chains in T_{AC}, the weight of node B's recommendations are effectively duplicated, and from Axiom 2, this duplication should not decrease the trust assessment. However for $T_{A'C'}$, no such duplication exists, and as such it's trust assessment should have a larger magnitude.

$$T_{AC} \geq T_{A'C'} \geq 0, |T_{A'C'} \geq 0 \tag{2.9}$$

$$T_{AC} \leq T_{A'C'} \leq 0, |T_{A'C'} \leq 0 \tag{2.10}$$

2.3.3 Topologies of Multi-Party Trust Networks

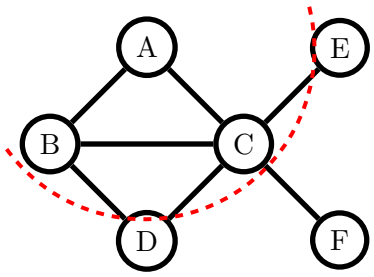
Beyond the attributes or characteristics of an individual trust relationship, within any multi party sparsely connected network or community, topological context is useful in both establishing trust and in disseminating observations for collaborative assessment.

Within sparsely connected networks, there are three primary types of relationship, minimally demonstrated in Fig. **??**;

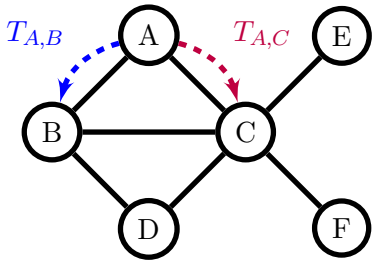
- *Direct* - Whereby two nodes have a 1-hop communications link between them (A,B,C in the given figure)
- *Indirect* - Where two nodes have a $n > 1$ hop communications link (E,D from A or Cs perspective in the given figure), i.e. there is no direct link from the trustor (A) to trustees (E,D)
- *Recommendation* - Where three nodes are fully connected so as to enable the exchange of direct opinions and form composite opinions based on the target and reporter (i.e. A has both its own Direct assessment of C, as well as it's knowledge of Bs Direct assessment of C)

2.3.4 Trust Establishment Strategies and their impact on Trust Frameworks

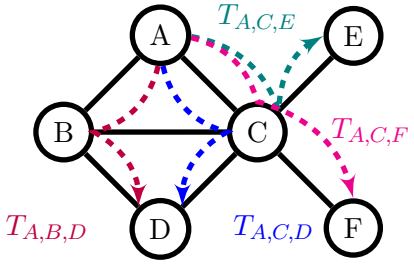
In the majority of cases, the establishment of trust is a purely observational effort, as opposed to providing a pre-initialised "secure state" [**? ? ?**]. The network is initialised, routes are dynamically generated and propogated depending on the routing strategy defined, and as information about the performance of nodes is recorded, Trust is established. This method of "blind trust establishment" may initially reduce efficiencies in the early phases of operation,



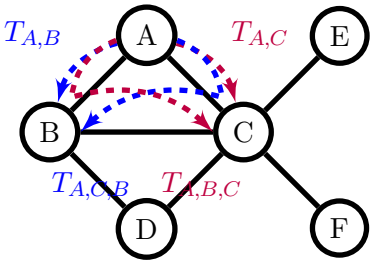
(a) Sample topology showing logical connections between nodes (Range of A shown in red dashed line)



(b) Direct Relationships, the two possible trust assessments from A to its connected neighbours, B,C



(c) Indirect Relationships, showing the four possible trust assessments from A or the three disconnected leaf nodes, D,E,F



(d) Recommender Relationship, showing the two discrete paths trust assessments travel to A; $T^R_{A,B} = T_{A,C} \cdot T_{C,B}$ and $T^R_{A,C} = T_{A,B} \cdot T_{B,C}$

Figure 2.10: Trust Topologies; Direct, Indirect, Recommender, etc. from the perspective of Node A

but this is quickly overcome with the establishment of a short "introduction" period, where the network is not tasked to perform any tasks where thorough Trust is necessary.

In the course of normal operations, there are two additional "trust establishment" cases that must be understood in the development of trust frameworks; newcomer integration and depart-return reintegration. In the first case, where a new node is joining the network, that new node is initially at a disadvantage; the rest of the network already trusts each other and could, by default, not perform any trusting behaviours to/with the new node, making it difficult for that new node to "demonstrate" trustworthiness [**?**]. This state is resolved by treating Trust as a multivariate factor consisting of at least two components; "Trust" and "Knowledge" or "Confidence", such that there is a substantive difference between "I'm not sure that this node is totally trustworthy but I've observed many actions" and "I have no idea how trustworthy this node is as I've not observed many actions". As such, when a node integrates with the trust network, the rest of the network is "ignorant" about it, and can probabilistically interact with it in a tentatively trusting manner.

In the latter operational case, where a node leaves the network (or "disappears" from the network in the case of temporarily broken topology such as in **??**) and returns, it is not sensible for nodes to maintain their previous opinions of the returning node. As such (and not purely for this reason) the majority of TMFs maintain a "remembering" (or forgetting) factor, normally noted as β. This *beta* factor inversely weights the contributions of older observations in the formation of

the current trust assessment, tending towards having zero confidence and zero trust in the node over time while it is not observed (or recommendations regarding that node are not observed).

On the return of a departed node, depending on the duration of the departure with respect to *beta*, it may be treated as a total newcomer, but will generally have some residual trust-valence, which puts it at a slight advantage compared to a complete newcomer. However, this reaction assumes that the identity of such a returning node can be verified, and in some cases, it is beneficial to initially distrust an apparrent "returner" in case of malicious masquerading.

2.3.5 Attacks on Trust

?] identify five types of attacks on Trust within networks that generate collaborative trust assessments through the exchange of recommendations; On-Off, Conflicting-behaviour, Badmouthing, Sybil and /Necomer attacks.

1. *Bad Mouthing:* Where a malicious node provides dishonest recommendations of other nodes in the network, to disrupt optimal operation by making other nodes appear untrustworthy

2. *Sybil attack:* Where a malicious node uses multiple pseudonymous entities to diffuse blame for bad behaviour, while maintaining a good reputation using the nodes genuine identity

3. *Newcomer Attack:* Similar to the Sybil attack in operation, whereby a malicious node will periodically assume a "fresh" identify with the network, shedding an identity that as accumulated an "untrustworthy" assessment

4. *Conflicting Behaviour attack:* Where a node (or a collaborating collection of malicious nodes) selectively drops messages, or modified recommendation values to or from certain nodes or groups of nodes, implicitly reducing the apparent trustworthiness of those attacked nodes with the rest of the fleet, while reducing the operational efficiency of the fleet overall. This attack exploits the dynamic properties of trust through taking advantage of the expected dynamism in topology and behaviors such that other nodes "won't notice" (Alternatively called the Grey Hole Attack)

5. *On-Off attack:* Malicious entities alternate between "good" and "bad" behaviours, hoping that they can remain undetected while causing damage. This attack exploits the dynamic properties of trust through time-domain inconsistent behaviours.

It is to be noted that the Sybil and Newcomer attacks do not exclusively rely on trust, but instead are a problem of network authentication.

All of these attacks can be abstracted as "non-isotropic attacks" i.e. attacks that attempt to hide malicious / selfish behaviour behind the expected statistical variation in observations within a cohort. In each case, a different dimension of this assumed statistical normality is exploited; in On-Off, the attacker attempts to "hide" in the time dimension by only occasionally misbehaving, in the Badmouth attack the attacker is relying on it's false recommendation being equitably received as its targets true actions. In the Conflicting behaviour attack, the attacker effectively "badmouths" a subset of nodes, hiding itself amid the "false" reports coming from the

conflicting subsets of nodes. Finally, in Sybil/Newcomer attacks the attacker takes advantage of an assumed naivety of the collective by presenting itself as a "new", and therefore, zero-history entity that can initially neither be trusted nor untrusted(See **??**)

2.4 Trusted Development and Operation of Autonomous Systems

2.4.1 Autonomy and Levels of Autonomy

Autonomy, like trust, is a nebulous term applied across research, defence and commercial circles that has its origins in human experience and interactions.

Autonomy, coming from the Greek roots *auto-* (self) and *nomos* (law) is the concept of a self-driven agency, and can be considered the concept of a "rational" individuals capacity to make un-coerced decisions in an informed manner. This autonomy is distinct from *freedom*, where freedom is the *ability* to perform an action, not the *capability to choose* which action to perform. That is not to say that autonomy or autonomous action exists in an ideal vacuum with perfect and complete information with no coercive factors or outside influences. The ability to recognise, process, weight and filter inputs, knowledge, "responsibilities", influences and outside factors and come to an effective decision is a key skill for any self-governing agent, however this is above and beyond the concept of "basic autonomy". From the implicit variability and complexity of environment and context that classically autonomous entities[3] inhabit, there is little assumption that "autonomy" always produces a categorically "correct"or "good" decision, but is instead a case of an agent choosing the action that is *in its own best interests based on available information*[**?**].[4]

This understanding of individual autonomy has been scaled up through social systems and has been studied at length to understand the emergence of post-Marxist proto-anarchistic movements [**?**] and from a higher perspective, international politics, especially in the cases of quasi-federalised collections of states such as the United States of America [**?**] and the European Union/Eurozone/Schengen Area [**?**]

In the most general case in the world of artificial systems, Autonomy is understood as a graduated spectrum of allocation of functionality between a system (or system of systems) and a human operator assigned with performing a given task. Where a system is more "autonomous", more of the sensing, planning, decision and action operations are performed by the system. (See Table **??** for a review of current definitions of autonomy and autonomous systems) This graduated spectrum of allocated functionality is generally termed the Level of Automation (LOA), where an increasing LOA correlated to increasing control and decision making freedom to the autonomous system

[3] That's *Homo Sapiens*

[4] Arply discusses a counter example of this "goodness" assessment as Huckleberry Finns' release of Jim against his "best judgement", and that rather than this action being an instance of morally justified or self-congratulatory autonomy, it was "the right thing to do" from an abstract moralistic perspective rather than a justifiably beneficial action, and it is a case of *akrasia*; the lacking of self-governance and the antonym of autonomy.

from the human operator(**??**). These levels can be loosely viewed as a spectrum from across Planning Support, Decision Support, Bounded Execution, and finally, Informed Execution[5].

While Autonomy is largely taken to be a robotics term based in the case of one human operator and one robotic entity, the development of more generalised cyber-physical systems has expanded this definition; from over-the-horizon human operation of Unmanned Aerial Vehicles (UAVs) to global networks of collaborating machines such as Google and beyond.

As such, the interactions *between* autonomous agents are becoming increasingly relevant to the operating efficiencies of overall collaborative systems, wether or not a human operator is "in-the-loop".

See **??** for a more thorough discussion on the Human Psychological Factors related to the planning, use, and integration of trusted autonomous systems in classical command and control contexts.

2.4.2 Trust Perspectives in Autonomous Operation

For the purposes of this work, two perspectives on trust for autonomous systems are defined: Design Trust and Operational Trust.

- *Design Trust* - When an autonomous system is under development a level of Trust is established in it through the manner in which it has been designed and tested. This is the same as conventional systems. Given that systems that have high-levels of autonomy are designed to behave adaptively to dynamic environments, it is challenging to fully predict such non-deterministic behaviours prior to operational deployment. For example, in a navigation system it is difficult to predict the dynamic environment it will need to adapt to. Trust needs to be developed so that the design and testing of such systems are sufficient to predict that operation will be, if not optimal, at least satisfactory.
- *Operational Trust* - Trust at runtime or in-situ that both the individual nodes within a system are operating as expected and that the interfaces between the operator and the system are as expected. This latter aspect covers issues such as physical/wireless links and interpretation of data at each end of such a communication link. This can be subdivided into two types of perspective;
  - *Hard Trust* or technical trust - The quantitative measurement and communication of the expectation of an actor performing a certain task, based on historic performance and through consensus building within a networked system. Can be thought of as a de-risking strategy to measure and monitor the ability of a system, or another actor within a system, to perform a task unsupervised.
  - *Soft Trust* or common trust - The qualitative assessment of the ability of an actor to perform a task or operation consistently and reliably based on social or experiential factors. This is the human form of trust and is the main motivational driver for the human-factors trust discussion in **??**. Can be viewed as the abstract level of confidence an operator has in an actor to perform a task unsupervised.

[5]In theory there is a further "Uninformed execution" level of autonomy, however this is beyond the scope of this work[**?**]

Table 2.6: Definitions of Autonomy

Definition	Source
...should be able to carry out its actions and to refine or modify the task and its own behaviour according to the current goal and execution context of its task	?]
Autonomy refers to systems capable of operating in the real-world environment without any for of external control for extended periods of time	?]
...a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect with it senses in the future. ...Exercises control over its own actions	?]
An unmanned systems own ability of sensing, perceiving, analysing, communicating, planning, decision-making, and acting, to achieve goals as assigned by its human operator(s) through designed HRI. ...The condition or quality of being self-governing	?]
...that the robot can operate self-contained, under all reasonable conditions without requiring recourse to the human operator. Autonomy means that a robot can adapt to change in its environment ...or itself ...and continue to reach a goal.	?]
...it should learn what it can to compensate for partial or incorrect prior knowledge	?]
Autonomy refers to a robot's ability to accommodate variations in its environment. Different robots exhibit different degrees of autonomy; the degree of autonomy is often measured by relating the degree at which the environment can be varied to the mean time between failures and other factors indicative of the robots performance.	?]
...agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal states.	?]
Systems have a set of intelligence based capabilities or learning adaptive capabilities that allow it to respond within a bounded domain to situations that were not pre-programmed or anticipated in the design.	?]

Hard Operational Trust is functionally derived from, but distinct from Design Trust.

It is already clear that these definitions are extremely close in their construction, but represent fundamentally different approaches to trust, one coming from a sociological perspective of person-to-person and person-to-group relationships from day to day life, and the other coming from a statistical or formal appraisal of an operation by a system during design, development, procurement, and deployment.

While the focus of this work is on the operational trust between teams of autonomous systems, there are two disjointed areas that are relevant to discuss. Firstly, it is valuable to briefly discuss the impact and effect of the actions and reactions of human operators in the context of autonomous operational trust; these human factors can have significant impact on the design constraints, operation, and indeed the ease-of-adoption of autonomous systems in the future.

Table 2.7: Levels of Decision Making Automation (Extended from ?])

LOA	Description
1	The computer offers no assistance; the human must make all decisions and actions
2	The computer offers a complete set of decision/action alternatives, or
3	Narrows the selection down to a few, or
4	Suggests one alternative and
5	Executes that suggestion if the human operator approves, or
6	Allows the human a restricted time to veto before automatic execution, or
7	Executes automatically, then necessarily informs the human, and
8	Informs the human only if asked, or
9	Informs the human only if it, the computer, decides to.
10	The computer decides everything and acts autonomously, ignoring the human.

Table 2.8: Levels of Automation (paraphrased from ?])

LOA	Description
Manual Control	The human monitors, generates options, selects options (makes decisions), and physically carries out options.
Action Support	The automation assists the human with execution of selected action. The human does perform some control actions.
Batch Processing	The human generates and selects options; then they are turned over to automation to be carried out (e.g., cruise control in automobiles)
Shared Control	Both the human and the automation generate possible decision options. The human has control of selecting which options to implement; however, carrying out the options is a shared task.
Decision Support	The automation generates decision options that the human can select. Once an option is selected, the automation implements it.
Blended Decision Making	The automation generates an option, selects it, and executes it if the human consents. The human may approve of the option selected by the automation, select another, or generate another option.
Rigid System	The automation provides a set of options and the human has to select one of them. Once selected, the automation carries out the function.
Supervisory Control	The automation selects and carries out an option. The human can have input in the alternatives generated by the automation.
Automated Decision Making	The automation generates options, selects, and carries out a desired option. The human monitors the system and intervenes if needed (in which case the level of automation becomes Decision Support).
Full Automation	The system carries out all actions.

Secondly, It is important to understand the wider Design Trust context to understand potential limitations or constraints on future development of such systems across their development lifecycle.

2.4.3 Summary of Human Factors impacting Operational Trust in Defence Contexts

When dealing with human supervision of autonomous or semi-autonomous systems, there is an inherent conflict between the expectations of the operator, and the hopes of system architects. System architects aim to provide more and more information to the operator to justify a systems operation, and Operators in reality need less and less information to be efficient when things are going well, and responsive in a dynamic environment. This places huge demands on Human Interface design and indeed on communications design to provide this timely, relevant, interactive connection between any autonomous system and the end operator(s). Recent work has presented the idea of taking user interface inspiration from the entertainment sector, in terms of UI best practises developed over two decades of Real-Time Strategy game development [**?**], and follow up work into automated mission debrief demonstrated that such operational support could improve causal situational awareness of an operator when compared to a human-baseline [**?**]. In terms of the human factors challenges [6], they are often contradictory in their direction, particularly when contrasting between Adaptive Automation and Cognitive Biases challenges. This is a key part of the "soft trust" perspective, where the operators and commanders need to be able to implicitly and explicitly trust the operation of a remote system with limited feed-back bandwidth, high latency, or long-term operation such that direct remote operation is infeasible or undesirable. To be able to trust that system's ability to continue on a course, survey an area, notify on detection of an anomaly, etc.is going to be the corner stone of any autonomous systems justification in the future.

2.4.4 Design Trust

As part of work conducted with DSTL [**?**], five aspects of Design Trust have been identifiedwith respect to Design Trust, with open research questions identified in each aspect emphasised.

1. **Formal Specification of Dynamic Operation**: Autonomous Systems (AS) may be required to operate in complex, uncertain environments and as such their specification may need to reflect an ability to deal with unspecified circumstances. This includes engaging with dynamic systems of systems environments where an autonomous system may cooperate with a system not envisaged at design time. *How can systems that are required to demonstrate that they meet their requirement be specified flexibly enough to permit adaptive behaviours*?

2. **Security**: Any unmanned system has the potential to be used for illegitimate purposes by unscrupulous third parties who could exploit security vulnerabilities to gain control of the system or sub-systems. Any system that has the potential to cause harm from such actions must have security designed in from the start to ensure that the system

[6] See **??** for a discussion of these challenges

can be trusted to be resilient from cyber attack. Current accreditation schemes rely on a security assessment of a known architecture and there are mutual accreditation recognition schemes that could be encoded in dynamic discovery handshake protocols. This would produce a secure network assured through the accreditation of its component systems. For example, the Multinational Security Accreditation Board (MSAB) deals with Combined Communications Electronics Board (CCEB) and NATO Accreditations to provide security assurance of internationally connected networks. Encoding such agreements into secure handshakes could enable dynamic accreditation of autonomous systems cooperating in a coalition environment. It is not known whether these have been demonstrated, so the question is: *Can autonomous systems be designed to understand the security situation when interfacing with known or unknown systems?*

3. **Verification and Validation of a Flexible Specification**: Following on from the description of a flexible specification, establish that the AS conforms and performs in accordance to the specification. This has direct implication for the trust in the resultant system. *How can systems demonstrate that they will behave acceptably when the environment is unknown?*

4. **Trust Modelling and Metrics**: This could be argued as part of the Verification and Validation of the system. However, models are increasingly being embedded into system design as a reference. Thus it is useful to consider this element separately. *How can trust be modelled sufficiently to span the space of most potential behaviours to help ensure that systems will be trusted when moved into operational environments? Can this be measured to allow comparison and minimum requirements set?*

5. **Certification**: The certification requirements placed on specific systems will vary depending on domain and national approaches to certification. However, the common element in the requirement for certification is that a certified system is deemed as sufficiently trustworthy for use within its context of certification. Additionally Certification also relies on the predictability of a system. Because the aim of autonomous systems is to deal effectively with uncertain environments, *can they (autonomous systems) be certified without being demonstrated in the environment within which they will adapt new behaviour?*

While this work is primarily concerned with those aspects of **Trust Modelling and Metrics**, it is useful to consider this Trust assessment in the wider context of the design process, and what best practices are available. Design against and Compliance with existing standards can contribute significantly to the demonstrable trustworthiness of any systems design. If a system has been designed to a Standard then it has known properties that have been accepted as good practice. However, current standards do not address the issue of the five areas listed above.

There are three main organisations that are developing or have developed assurance standards for Unmanned Systems in commercial, civil and military applications:

- NATO Standardization Office (NSO)
- Society of Automotive Engineers (SAE)
- American Society of Testing and Materials (ASTM)

LOI	Description
1	Indirect receipt/transmission of UAV related payload data
2	Direct receipt of ISR data where direct covers reception of UAV payload data by the UCS when it has direct communication with the UAV
3	Control and monitoring of the UAV payload in addition to direct receipt of ISR/other data
4	Control and monitoring of the UAV, less launch and recovery
5	Launch and Recovery in addition to LOI 4

Table 2.9: Levels of Interoperability for STANAG 4586 Compliant UCS [**?**]

NATO Standardization Office Faced with the growing adoption of similar but disparate UAV systems within NATO territories and coalition nations, STANAG 4586[**?**] was promulgated in 2005 and defined a logistic and interoperability framework to provide commonality in the command and control architecture and implementations of UAV/Ground station communications.

This included a particularly interesting development in the form of Society of Automotive Engineers (SAE) Vehicle Specific Module (VSM) interoperability, whereby existing systems could be grandfathered into STANAG 4586 compliance by the addition of a VSM to operate as a protocol translator. This VSM could be mounted on the remote system directly, utilising a compliant Data Link Interface (DLI), or mounted on the ground-based controller, retaining the proprietary DLI to the remote system. The standard describes five Level of Interoperabilty (LOI) for compliant UAV systems, shown in Table **??**. This structure has been criticised as being short sighted and at odds with the reality of modern and proposed autonomous vehicle operations [**?**], specifically that in modern autonomous systems, there is no such thing as "direct control" or "Operator-in-the-loop", especially in the case of Beyond Line of Sight (BLOS) systems, and that in increasingly autonomous systems, operation is done as Human Supervisory Control (HSC), or more commonly described as Operator-on-the-loop, whereby the operator interacts with the intermediate autonomous system and that autonomous system eventually performs that task on the hardware.

Further, the standard predominantly deals with a one-to-one mapping between operators and nodes, when this is quite against the current state of the art; greater focus is being made in collective and collaborative assignment and having a single operating agent managing groups of autonomous nodes in-field, and handing off vehicle management responsibilities to the individual nodes.

SAE The AS-4 steering group is responsible for the development and maintenance of the Joint Architecture for Unmanned Systems (JAUS) standards, which provide several service sets for Inter-System cooperation and interoperability, either in the form of a specified design language (JSIDL[7]) or as a direct framework implementation, such as the JAUS Mobility, Mission Spooling,

[7] JAUS Service Interface Definition Language

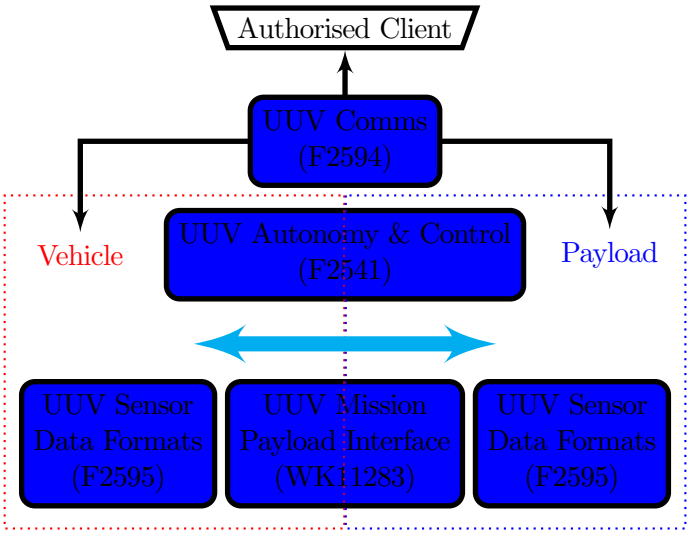


Figure 2.11: ASTM F41 UMVS Architecture (with relevant substandards in parenthesis)

Environment Sensing, or Manipulator Service Sets[8]. This provides a stack-like interoperability model akin to the OSI inter-networking standard, providing logical connections between common levels across devices regardless of how subordinate layers are implemented. Importantly, JAUS service models are open-sourced under the BSD-license, and a development toolkit is available for anyone to develop JAUS-compatible communications and control protocols[**?**].

It is also important to note that JAUS is part funded, and heavily utilised by, US Army and Marine Robotic Systems Joint Project Office (RS-JPO), which manage the development, testing, and fielding of unmanned (ground) systems for those respective forces.[**? ?**] This includes now legacy M160 mine clearance platform and the highly popular (both with forces and their in-field operators) iRobot Packbot inspection and Explosive Ordnance Disposal (EOD) family of robotic platforms.

American Society of Testing and Materials (ASTM) The ASTM F38 committee has developed a Line of Sight (LOS), single-asset-single-operator stove-piped framework for Unmanned Air Systems that is too constrained in scope for applicability to a more heterogeneous operating environment[**?**]. However, the F41 Committee, focused on Unmanned Maritime Vehicle Systems (UMVSs) has collectively developed a range of interoperable standards, covering Communications, Autonomy and Control, Sensor Data Formats, and Mission Payload Interfacing. Of particular interest is the Autonomy and Control standard which highlighted a requirement on the vehicle system to be able to recognise an authorised client, be that a human operator or an additional collaborating vehicle [**?**]. Further, the standard states that the responsibility of the safety and integrity of any payload remains with the vehicle. This standard was withdrawn in 2015 due to ASTM regulations requiring standards to be updated within 8 years of approval, and has no direct replacement within ASTM, but stands as a useful guiding perspective on autonomy standards within industry.

[8]SAE AS6009, AS 6062, AS 6060, and AS 6057 respectively

Summary of Design Trust

The implications of trust in autonomy beyond securing communications and data are an area in need of further research. Of particular concern is the verification of autonomous behaviours and failsafe behaviour [**?**]. The addition of increased on-board autonomy in MUxS, properly understood and verified, would greatly improve this future capability, similar to recent developments in the UMVS arena [**?**].

There are opportunities for increased decentralisation and in-field collaboration [**?**], however, difficulties in "Trust" between human operators and autonomous systems have already been clearly identified [**?**],and this has been demonstrated by the recent decision by the German government to renege on its €500M investment in the Euro Hawk programme, due to concerns about civil certification of the onboard autonomy [**?**] In order for these new distributed structures to be relied upon to provide operational performance, reliability and to maintain in-field situational awareness, vulnerabilities to disruption, interruption, and subversion need to be understood and minimised.

2.5 Trust in Autonomous MANETs

2.5.1 Trust Model Design Considerations

From the previous sections, Trust can be redefined as "the level of confidence one agent has in another to perform a given action on request or in a certain context". Trust in the autonomous or semi-autonomous realm is the ability of a system to establish and maintain this level of confidence in itself or another systems' operations.

There are five topics that are important to address in any MANETs trust model [**?**]:

- The trust model should be without infrastructure. Because the network routing infrastructure is formed in an ad-hoc fashion, the trust management can not depend on, e.g., a Trusted Third Party (TTP). There is no PKI, where some center nodes monitor the network, and publish illegal nodes periodically. In a MANET, there are no certification authorities (CA) or registration authorities (RA) with elevated privileges etc.
- The trust model should be anonymous because of the anonymity of mobile nodes in MANETs.
- The trust model should be robust. That is, it can be robust to all kinds of unfriendly attacks and the network itself should not be susceptible to attacks by unfriendly nodes. Moreover, in the presence of malicious nodes, they may attempt to subvert the model in order to get an unfairly good trust value.
- The trust model should have minimal control overhead in accordance with computation, storage, and complexity.
- The trust model should be self-organized. MANETs are characterized to have dynamic, random, rapidly changing and multi-hop topologies composed of variably bandwidth-constrained links

2.5.2 Vulnerabilities of MANETs

The openness of the MANET architecture leaves it inherently vulnerable to security threats. Mitigation protocols must be built on top of this architecture to maintain security and reliability in the face of open-access threats (whether these threats be directed attacks, uncooperative or selfish operation, or indeed malfunctioning/failing nodes). It is worthwhile to briefly summarise the factors of MANET architecture that make it vulnerable to different threat vectors, establish the inter-node threat surface within an operating MANET.

Exposed Threat Surfaces

This section based on a summary of **?** *] and other sources directly cited*

Wireless Links - The use of wireless interfaces expose such networks to eavesdropping and active interference, with no physical access required. These links normally have significant channel access and bandwidth constraints compared to closed-wired networks. This presents outside threats with the ability to eavesdrop or actively interfere with the operation of the network, leading to data security risk and risk of DoS-style attacks.

Mobility and Dynamic Topology - Nodes joining/leaving/re-joining the network and moving around the environment leads to significant topology and access control changes, making it difficult to differentiate between malicious and normal behaviour. Additionally this assumption of node mobility and "temporary disconnection" presents opportunities to outside attackers to physically compromise, capture or replicate nodes.

Assumption of Cooperation - MANET routing is predicated on the assumed "fairness" of nodes both in their routing operation and their "advertisement" of routing capability, leading to opportunities to disrupt the optimal operation of the network [**?**].

Resource Constraint - In MANETs, more than in static wired networks, secondary resources such as power, and tertiary resources such as locomotion, onboard processing and data storage capabilities present additional opportunities for selfish or malicious threat that simultaneously constrain the ability of nodes to mitigate threat (i.e. limited processing power restricting the use of advanced cryptographic protocols, power/locomotion constraints limiting the available operational time to "learn" about attack characteristics, etc.)

Insecure/Fuzzy Operational Boundary - With no hard boundary between "in network" and "out of network", MANET security must combat both internal and external threats.

Threat Mitigation Strategies

Many classical mitigation strategies focus on selfishness rather than malicious attack, usually including some form of misbehaviour-induced backoff policy to passively punish misbehaviour [**?** **?**]. On the other hand, many strategies are usually based on some derivative form of hardline-network Intrusion Detection System (IDS) frameworks, where such systems passively observe a network for misbehaviour and

?] first proposed a general IDS framework for MANETs; leveraging the distributed and cooperative predicates of the architecture, introducing both per-node and cooperative IDS

Table 2.10: Selected Attacks on the Protocol Stack extended from **?**]

Layer	Attacks
Application	Data Corruption, Malware, Virii and Worms
Transport	SYN Flodding, Session Hijacking
Network	"HELLO" Flood, (Black/Worm/Sink)hole, Sybil, Replay, Rishing, Resource-Consumption
Data Link	Monitoring, Traffic Analysis
Physical	Eavesdropping, Active Interference

Table 2.11: Threat Actor Classification from **?**]

Emission	Location	Quantity	Target	Rationality	Mobility
Active	Insider	Individual	Confidentiality	Rational	Static
Passive	Outsider	Collaborating	Integrity	Irrational	Mobile
			Fairness		
			Authorisation		
			DoS		

submodules such that nodes work independently and cooperatively to identify certain misbehaviours, specifically targeting routing attacks and incongruities. Extensions to this framework were also proposed to have a range of these modules operate at each layer of the protocol stack to improve detection response and range [**?**] However it is commonly accepted that these frameworks had significant deficiencies in terms of power and communications overheads [**? ?**]. One interesting aspect to highlight about **?**] is that while information about the physical mobility was assessed in the detection of misbehaviours in the context of routing changes, no collective cross-comparison was used, and rather focused on a per-node relative distance/velocity estimation to identify anomalous routing table updates.

2.5.3 Trust Management Frameworks

Distributed trust management frameworks for MANETs aim to detect, identify, and mitigate the impacts of malicious or selfish actors by generating, distributing and integrating per-node assessments and opinions to collectively self-police behaviour. From the settled upon definition of trust (From **??**), these opinions are attempting to model the confidence of success in a particular actor for a particular future action.

This predictive behaviour attempts to solve four important problems (paraphrased from [**?**]):

- *Decision support* - For example; making informed routing table decisions based on past successes/failures.
- *Adaptability* - Ongoing prediction of the networks future trust states directly determines the risk faced by the network. Internalised knowledge of the expected risk can aid in

selecting appropriate measures/ countermeasures such as automatically varying the level of authentication required for network activities.

- *Misbehaviour Detection* - Trust evaluation leads to a the natural policy that highly variable or low-trust nodes within a network should be subject to higher scrutiny; triggering this response indicates that a node is damaged or misbehaving.
- *Abstraction of Collective security characteristics* - Through per-node trust evaluation, the generalised trustworthiness of a set or subset of nodes can be derived to encapsulate the "health" of the network as a whole.

Various models and algorithms for describing trust and developing trust management in distributed systems, Peer to Peer (P2P) communities or wireless networks have been considered. Taking some examples;

- *Hermes Trust Establishment Framework* uses a Bayesian Beta function to model per-link PLR over time, combining "Trust" and "Confidence of Assessment" into a single value [**?**].
- *Objective Trust Management Framework (OTMF)* takes a Bayesian approach and introduces the idea of applying a Beta function to changes in the per-link PLR over time, combining "Trust" and "Confidence of Assessment" into a single value [**?**]. OTMF however does not appropriately combat multi-node-collusion in the network [**?**].
- *Trust-based Secure Routing* demonstrated an extension to DSR, incorporating a Hidden Markov Model of the wider ad-hoc network, reducing the efficacy of Byzantine attacks, particularly black-hole attacks but is limited by focusing on single metric observation (PLR) [**? ?**].
- *Cooperation Of Nodes: Fairness In Dynamic Adhoc NeTworks (CONFIDANT)*; presented an approach using a probabilistic estimation of normal observations, similar to OTMF. Also introduced a greedy topology weighting scheme that internally weighted incoming trust assessments based on historical experience of the reporter [**?**].
- *Fuzzy Trust-Based Filtering*;presented a method using Fuzzy Inference to cope with imperfect or malicious recommendation based on a probabilistic estimation of performance using conditional similarity to classify performance using overlapping Fuzzy Set Membership functions to collaboratively filter reputations across a network [**?**].
- *Multi-parameter Trust Framework for MANETs (MTFM)* uses a number of communications metrics together for form a vector of trust, apply grey information theory to allow a system to detect and identify the tactics being used to undermine or subvert trust [**?**].

2.5.4 Single Metric Trust Frameworks

The Hermes trust establishment framework [**?**] uses Bayesian reasoning to generate a posterior distribution function of "belief", or trust, given a sequence of observations of that behaviour, $p(B|O)$(**??**).

$$p(B|O) = \frac{p(O|B) \times p(B)}{\rho} \tag{2.11}$$

Where $p(B)$ is the prior probability density function for the expected normal behaviour, and ρ is a normalising factor.

Due to it's flexibility and simplicity, Hermes assumes that $p(B)$ is a Beta function ((**??**)), and therefore the evaluation of this trust assessment is based around the expectation value of the distribution (**??**) where α and β represent the number of successful and unsuccessful interactions respectively for a particular node i.

$$\text{beta}(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} \tag{2.12}$$

$$E(p) = \frac{\alpha}{\alpha+\beta} \tag{2.13}$$

$$\text{where } 0 \leq p \leq 1; \alpha,\beta > 0$$

A secondary measurement of the confidence factor of the trust assessment t is generated as (**??**) and these measurements are combined to form a "trustworthiness" value T (**??**).

$$t_i \rightarrow E[\text{beta}(p|\alpha,\beta)] = \frac{\alpha_i}{\alpha_i+\beta_i} \tag{2.14}$$

$$c_i = 1 - \sqrt{\frac{12\alpha_i\beta_i}{(\alpha_i+\beta_i)^2(\alpha_i+\beta_i+1)}} \tag{2.15}$$

$$T_i = 1 - \frac{\sqrt{\frac{(t_i-1)^2}{x^2} + \frac{(c_i-1)^2}{y^2}}}{\sqrt{\frac{1}{x^2}+\frac{1}{y^2}}} \tag{2.16}$$

In (**??**), x and y are constants to weight the two-dimensional polar mapping of trust and confidence assessments (t_i,c_i) From **?**] these are set to $x=\sqrt{2}, y=\sqrt{9}$, generating an elliptical mapping $f(t,c) \mapsto T$, effectively weighting the level of confidence heavier than the observed trust behaviour.

Upon this per-node assessment methodology, OTMF overlays an observation distribution protocol so as to make the measurements α_i and β_i representative of the direct and 1-hop networks observations of the target node i, as well as expiring old observations from assessment and eliminating observations from "untrustworthy" nodes.

To date this work has been mostly limited to terrestrial, RF based networks. There are many situations where the observed metrics will include significant noise and occur at irregular, sparse, intervals. Conventional approaches such as probabilistic estimation do not produce trust values that reflect the underlying reality and context of the metrics available, as they require a-priori assumption that the trust value under exploration has an expected distribution, that that distribution is mono-modal, and the input metrics are binary. In scenarios with variable, sparse, noisy metrics, estimating the distribution is difficult to accomplish a-priori. These single metric TMFs provide malicious actors with a significant advantage if their activity is undetectable by that one assessed metric, especially if the attacker is aware of the observed metric in advance.

The objective of operating a TMF is to increase the confidence in, and efficiency of, a system by reducing the amount of undetectable negative operations an attacker can perform. In the case where the attacker can subvert the TMF, the metric under assessment by that TMF does not cover the threat mounted by the attacker. In turn, this causes a super-linearly negative

effect in the efficiency of the network as the TMF is assumed to have reduced the possible set of attacks when in fact it has only made it more advantageous to attack a different aspect of the networks operation. An example of such a behaviour would be the case in a TMF focused on PLR where an attacker selectively delays packets going through it, reducing the overall throughput of one or more network routes. Such behaviour would not be detected by the TMF.

2.5.5 Multi-Metric Trust Frameworks

Given the potential incentives to a selfish attacker and potential threats to trust and fairness in sparse, noisy, and constrained environments, single metric trusts discussed above do not suitably cover the exposed threat surface.

A multi-metric approach may be more appropriate to capture and monitor the realities of harsh and sparse communications environments.

MTFM [**?**] uses Grey Theory (see **??**) to perform cohort based normalization of metrics at runtime, providing a "grey relational grade" of trust compared to other observed nodes in that interval for individual metrics, while maintaining the ability to reduce trust values down to a stable assessment range for decision support without requiring every environment entered into to be characterised. This presents a stark difference between the Grey and Probabilistic approaches. Grey assessments are relative in both fairly and unfairly operating networks. All nodes will receive mid-range trust assessments if there are no malicious actors as there is nothing "bad" to compare against, and variations in assessment will be primarily driven by topological and environmental factors. **?**] demonstrated the ability of Grey Relational Analysis (GRC) to normalise and combine disparate traits of a communications link such as instantaneous throughput/load, received signal strength, etc. into a Grey Relational Coefficient (GRC), or a "trust vector" in this instance.

The grey relational vector is given as

$$\theta^t_{k,j} = \frac{\min_k|a^t_{k,j}-g^t_j|+\rho\max_k|a^t_{k,j}-g^t_j|}{|a^t_{k,j}-g^t_j|+\rho\max_k|a^t_{k,j}-g^t_j|}$$
$$\phi^t_{k,j} = \frac{\min_k|a^t_{k,j}-b^t_j|+\rho\max_k|a^t_{k,j}-b^t_j|}{|a^t_{k,j}-b^t_j|+\rho\max_k|a^t_{k,j}-b^t_j|} \tag{2.17}$$

where $a^t_{k,j}$ is the value of an observed metric x_j for a given node k at time t, ρ is a distinguishing coefficient set to 0.5, g and b are respectively the "good" and "bad" reference metric sequences from $\{a^t_{k,j}k=1,2...K\}$, i.e. $g_j=\max_k(a^t_{k,j})$, $b_j=\min_k(a^t_{k,j})$ (where each metric is selected to be monotonically positive for trust assessment, e.g. higher throughput is presumed to be always better).

Weighting can be applied before generating a scalar value (**??**) allowing the detection and classification of misbehaviours.

$$[\theta^t_k,\phi^t_k] = \left[\sum_{j=0}^{M} h_j\theta^t_{k,j}, \sum_{j=0}^{M} h_j\phi^t_{k,j}\right] \tag{2.18}$$

Where $H=[h_0...h_M]$ is a metric weighting vector such that $\sum h_j=1$, and in unweighted case, $H=[\frac{1}{M},\frac{1}{M}...\frac{1}{M}]$. θ and ϕ are then scaled to [0,1] using the mapping $y=1.5x-0.5$. To minimise

the uncertainties of belonging to either best (g) or worst (b) sequences in (**??**) the $[\theta,\phi]$ values are reduced into a scalar trust value by $T_k^t = (1+(\phi_k^t)^2/(\theta_k^t)^2)^{-1}$ [**?**]. MTFM combines this GRC with a topology-aware weighting scheme (**??**) and a fuzzy whitenization model (**??**). This whitenization model allows the previously grey value to be practically computed on and used to generate the final trust assessment, through fuzzy sequence classification, functionally treating each observation differently if it appear to be Trusted, Untrusted or Ignorant (**??**) [**?**]. See **??** for a wider discussion of this topic and the operation of Grey numbers.

There are three classes of topological trust relationship used; Direct, Recommendation, and Indirect, as discussed in **??**. Where an observing node n_i assesses the trust of another target node, n_j; the Direct relationship is n_i's own observations n_j's behaviour. In the Recommendation case, a node n_k which shares Direct relationships with both n_i and n_j, gives its assessment of n_j to n_i. In the Indirect case, similar to the Recommendation case, the recommender n_k does not have a direct link with the observer n_i but n_k has a Direct link with the target node, n_j. These relationships give node sets, N_R and N_I containing the nodes that have recommendation or indirect, relationships to the observing node respectively.

$$\begin{aligned} T_{i,j}^{MTFM} = & \frac{1}{2} \cdot \max_s \{f_s(T_{i,j})\} T_{i,j} \\ & + \frac{1}{2} \frac{2|N_R|}{2|N_R|+|N_I|} \sum_{n \in N_R} \max_s \{f_s(T_{i,n})\} T_{i,n} \\ & + \frac{1}{2} \frac{|N_I|}{2|N_R|+|N_I|} \sum_{n \in N_I} \max_s \{f_s(T_{i,n})\} T_{i,n} \end{aligned} \tag{2.19}$$

Where $T_{i,n}$ is the subjective trust assessment of n_i by n_n, and $f_s = [f_1, f_2, f_3]$ given as:

$$\begin{aligned} f_1(x) &= -x+1 \\ f_2(x) &= \begin{cases} 2x & \text{if } x \leq 0.5 \\ -2x+2 & \text{if } x > 0.5 \end{cases} \\ f_3(x) &= x \end{aligned} \tag{2.20}$$

In the case of the terrestrial communications network used in [**?**], the observed metric set $X = x_1,...,x_M$ representing the measurements taken by each node of its neighbours at least interval, is defined as $X =$ [packet loss rate, signal strength, data rate, delay, throughput].

?] demonstrated that when compared against OTMF and Hermes trust assessment, MTFM provided increased variation in trust assessment over time, providing more information about the nodes' behaviours than packet delivery probability alone can.

2.6 Conclusion

In this chapter, MANET implementations, topologies, and applications have been explored. Further, the concept of a "Trusting" network has been explored, both on a abstract theoretical basis, and in the context of a wider development and operational pipeline involving autonomous

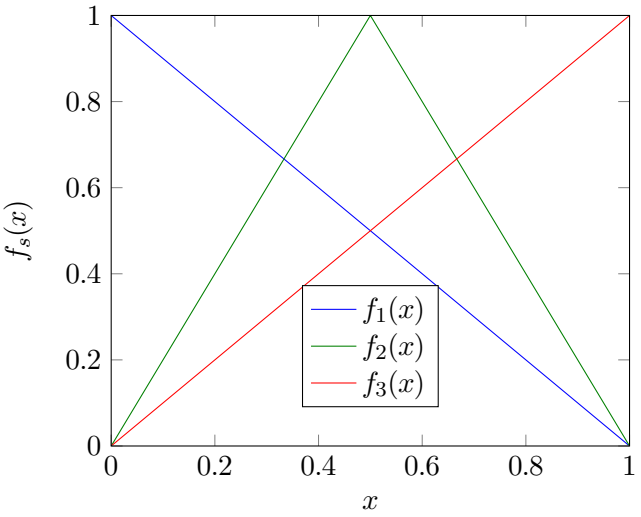


Figure 2.12: Operation of the Centre-Point Triangular Whitenization used in MTFM **??**

actors, including a review of current Trust Management Frameworks (TMFs). These TMFs have several aspects that are dependant on assumptions of sufficient available resources and connectivity, that they may not behave as efficiently as would be hoped in constrained or delayful networks, particularly in cases where metric assessment data may be extremely sparse or noisy.

While many aspects of this Trust have been discussed, we are primarily concerned with this problem of assessment of trust through experiential observation of node behaviours in a practical runtime environment, namely the underwater acoustic environment. In the next chapter, the marine communications environment will be studied, as will the current state of the art in the use of autonomy in defence related maritime applications, and briefly discussing the context of those operations.

Chapter 3

Maritime Communications and Operations

3.1 Maritime Communications Environment

The key challenges of underwater acoustic communications are centred around the impact of slow and differential propagation of energy (RF, Optical, Acoustic) through water, and it's interfaces with the seabed / air. The resultant challenges include; long delays due to propagation, significant inter-symbol interference and Doppler spreading, fast and slow fading due to environmental effects (aquatic flora/fauna; surface weather), carrier-frequency dependent signal attenuation, multipath caused by the medium interfaces at the surface and seabed, variations in propagation speed due to depth dependant effects (salinity, temperature, pressure, gaseous concentrations and bubbling), and subsequent refractive spreading and lensing due to that same propagation variation[**?**].

3.1.1 Mechanics of Acoustic Transmission

Unlike in RF energy transfer (where photons move through space to transmit energy from one place to another), acoustic waves are the result of mechanical perturbation of a medium where localised compressions and extensions pass energy across a medium through that mediums elastic properties. These "compression waves" propagate away from its source, and the rate of this propagation is the sound speed, velocity or c, measured in ms^{-1}. This is not to be confused with the fluid velocity corresponding to the instantaneous motion of particles in the medium.

Hydrophones, like their more common microphone equivalent in air, are fundamentally pressure sensors. Acoustic pressure is usually measured in *Pascals* ($Pa/\mu Pa$). In the underwater environment, the dynamic range (difference between instantaneous high and low pressure values) may be extremely high, often more than 10 orders of magnitude higher. As such, logarithmic notation is justified.

Useful acoustic signals are generally maintained vibrations rather than instantaneous pulses. They are characterised by their frequency f expressed in Hertz (Hz) or by their Period (T) in seconds. In commonly used underwater acoustics, used frequencies range from $\approx 10Hz - 100kHz$ depending on application [**?**].

Table 3.1: Summary of physical factors differentiating terrestrial and acoustic channel constraints

Variable	Air	Water
Density	1.3kgm^{-3}	$\approx 1027\text{kgm}^{-3} \pm 0.5\%$
Speed of Sound	340ms^{-1}	$\approx 1500\text{ms}^{-1} \pm 5\%$
Speed of Light	$2.99 \times 10^8 \text{ms}^{-1}$	$2.249 \times 10^8 \text{ms}^{-1}$

As with all waves, the relationship between frequency, period and the wavelength is given as in (**??**). As such the generally used upper and lower bounds of wavelength in most applications is from 1.5m@10Hz to 0.015m@100kHz.

$$\lambda = cT = \frac{c}{f} \tag{3.1}$$

This wide range of frequencies and wavelengths allow for a diverse set of constraining factors; (Paraphrased from **?**]).

- *Attenuation* in water; limiting the maximum usable range, which increases very rapidly with frequency
- *Dimensions* of sound source; which increase at lower f for a given transmission power
- *Spatial Selectivity* of sources and receivers as f increases, due to similarly increasing directivity of energy propagation.
- *Acoustic Response* of target surfaces (analogous to receiver gain in RF networks).

3.1.2 Velocity and density

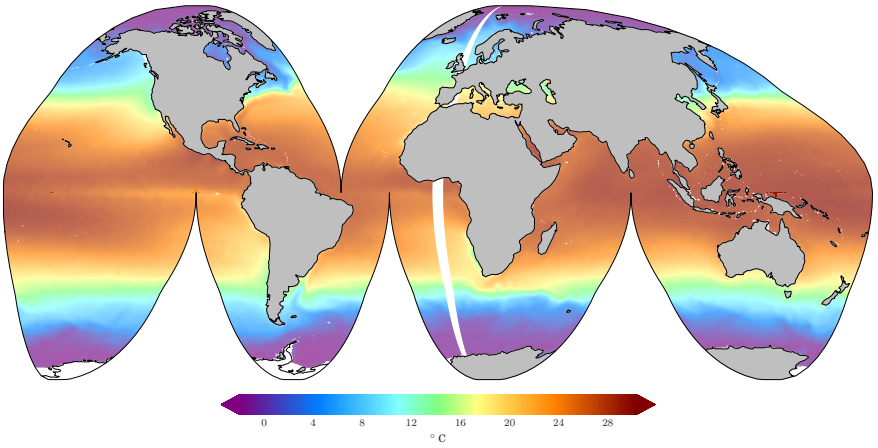
Air has a baseline density of approximately 1.3kgm^{-3}, and the speed of sound is typically static around 340ms^{-1}. In sea water, acoustic wave velocity is close to $c = 1500\text{ms}^{-1}$ (generally between $1450 to 1550\text{ms}^{-1}$ depending on temperature, pressure, salinity etc.). Similarly variable is sea water density, which is nominally $\rho = 1027\text{kgm}^{-3}$[**?**].

While the sea/air surface is (ideally) a simple refractive interface, the interface between open seawater and marine sediment is graduated, with density ranges between $1200 to 2000\text{kgm}^{-3}$. This results in refractive and reflective velocities in the sediment interface ranging from $1500 to 2000\text{ms}^{-1}$[**?**].

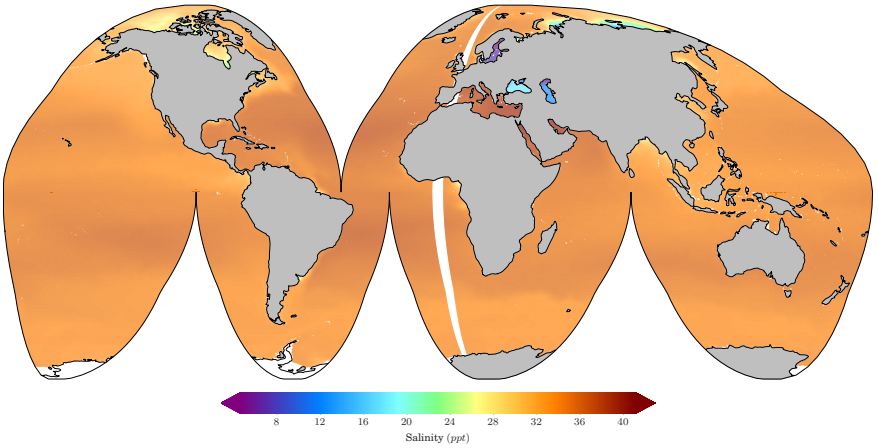
For comparison, the speed of light in air/water is $2.99 \times 10^8\,\text{ms}^{-1}$ and $2.249 \times 10^8\,\text{ms}^{-1}$ respectively.

?] proposed a more accurate model of acoustic velocity incorporating archival data from 15 worldwide sites that takes Temperature, Salinity and Depth into consideration.

$$\begin{aligned} c = {} & 1448.96 + 4.591T - 5.304 \times 10^{-2} T^2 + 2.374 \times 10^{-4} T^3 \\ & + 1.340(S-35) + 1.630 \times 10^{-2} D + 1.675 \times 10^{-7} D^2 \\ & - 1.025 \times 10^{-2} T(S-25) - 7.139 \times 10^{-13} T D^3 \end{aligned} \tag{3.2}$$



(a) Surface Temperature



(b) Surface Salinity

Figure 3.1: Global Variations in selected Speed of Sound variance factors

Where T is the temperature in Celsius, S the salinity in parts per thousand, and D is the depth below the surface in meters.

These are "ideal" assessments, and the parameters of this model are massively varied both across individual water columns, and indeed across bodies of water, across the world. National Oceanographic and Atmospheric Association (NOAA) regularly publish their World Ocean Atlas (WOA) [**?** **?**] that includes these variables and many more, sampled across the world and at a range of depths. Outputs from these surveys are shown for example in **??**. Variability in Temperature across the globe is something that we are acutely aware of, but the significant regional variations, such as in the Sea of Japan, the US Eastern Seaboard, and between the Mediterranean and Black Seas (**??**). Global surface salinity appears almost uniform in comparison (**??**). However, a few global variants stand out, both due to the extremity of their transition in relatively small areas, and the general research / defence context of those areas. For example the differential between the geographically proximate and politically contentious Black, Caspian, and Mediterranean Seas, as well as the Persian Gulf exhibit variations from less than 6*ppt* to over 40*ppt*. Similarly, there is a navigable waterway providing access between

the Baltic and North seas that across the 300km long run from Malmö in Sweden to Skagen in Denmark, transitions from less than 5*ppt* to just under 30*ppt*.

Below the surface, the variability increases; **??** shows an example of a depth profile of these variations and the modelled impact on the speed of sound with respect to depth in three different regions. The variability of this speed is crucial to the operation of an underwater acoustic network, as it fundamentally changes the propagation paths of compressive energy transfer, and in particular, the fastest "path". **??** shows the impact on this fastest recieved path and it's true path between two nodes in shallow, littoral, waters. Even with relatively small variations in sound speed, and with the introduction of sea floor/surface interfaces, the "fastest" path deviates significantly from a true "line of sight" path where there is any variability in speed of sound profile, making delay-based positioning extremely difficult, and presents significant opportunities for out-of-sync multi-path effects[1]

3.1.3 Intensity and Power

The energy of an acoustic wave is encapsulated into its kinetic and potential parts; where its kinetic energy corresponds to the active motion energy of the particles in the medium, and the potential energy corresponding to the elastic potential of the medium in displacement/compression.

The acoustic intensity (I) is the energy flux mean value per unit of surface and time (**??**) in Watts/m^2 where p_0 is the plane wave amplitude (pressure) and $P_{rms} = p_0/\sqrt{2}$

$$I = \frac{p_0^2}{2\rho c} = \frac{p_{rms}^2}{\rho c} \tag{3.3}$$

3.1.4 Attenuation

The attenuation that occurs in an underwater acoustic channel over a distance d for a signal about frequency f in linear (**??**) and dB forms (**??**) is given as;

$$A_{\text{aco}}(d,f) = A_0 d^k a(f)^d \tag{3.4}$$

$$10\log A_{\text{aco}}(d,f)/A_0 = k \cdot 10\log d + d \cdot 10\log a(f) \tag{3.5}$$

where A_0 is a unit-normalising constant, k is a geometric spreading factor (commonly taken as 1.5 for practical use, but may be 2 for perfect spherical propagation or 1 for perfect plane-wave propagation)), and $a(f)$ is the absorption coefficient, that may be modelled in a variety of ways.

Thorp's formula (**??**) is very simple, only depending on f, and is designed to be most accurate about a temperature of 4°C at a depth of $\approx 1Km$. The Ainslie & McColm model is more complex, and incorporates the acidity of the water (H^+) as well as temperature (T), salinity (S in parts per trillion) but not depth (**??**). The Fisher-Simmons model (**??**) is significantly more complex, taking into account the effects of boric acid concentrations and dissolved magnesium

[1] **??** shows a staged, iterative approximation method to arrive at the shortest path, where the "colour intensity" of the chart shows the stage at which that path was explored, so the "final" paths are darkest, and the "exploitative" paths are lightest, however in reality these secondary paths are still emitted and arrive, delayed, to the reciever, causing significant inter-symbol interference unless equalised.

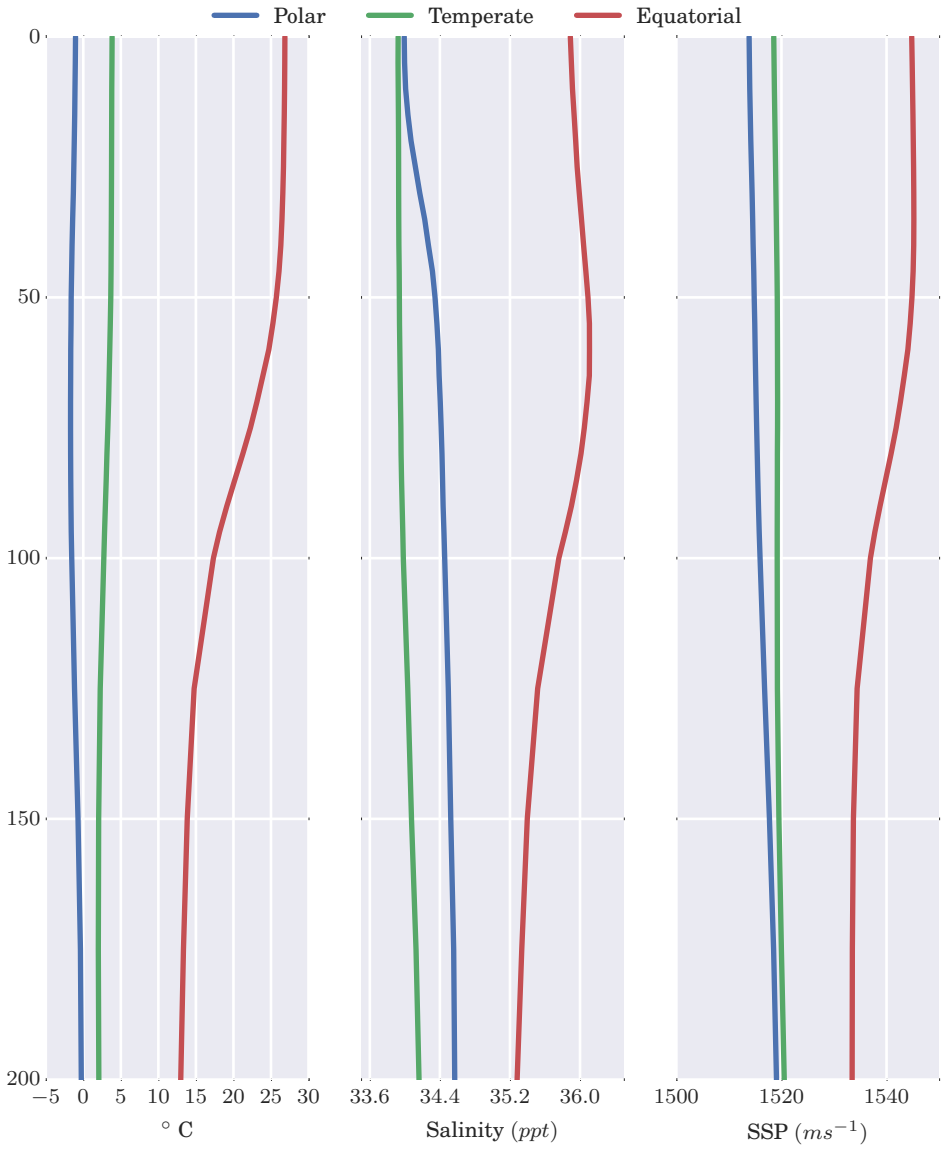
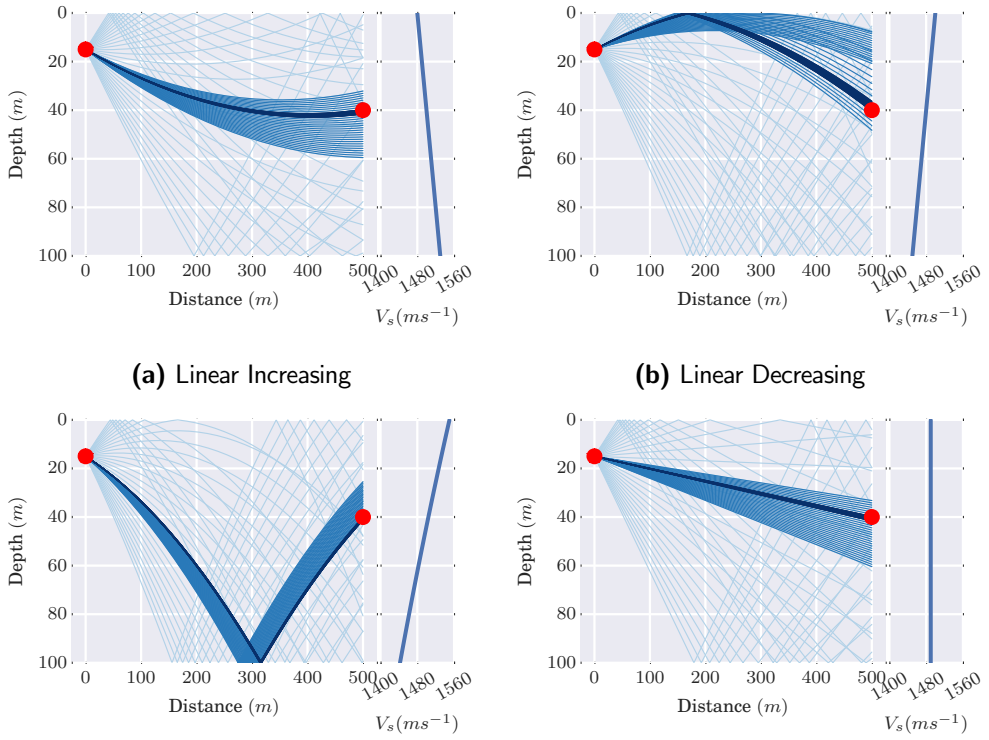


Figure 3.2: Depth Variations in selected Speed of Sound variance factors

sulphate. While there are several limitations on this model in terms of its being fixed at a salinity of 35 ppt and a pH of 8, as this model incorporates depth, temperature, distance and frequency, it is very attractive for research directed at high variability environments and is used for the remainder of this work unless otherwise stated.

Regardless of the variations of particular attenuation models, comparing $A_{aco}(d,f)$ with the RF Free-Space Path Loss model **??**, the impact of range on signal power is exponential underwater, rather than quadratic in terrestrial RF ($A_{\text{aco}} \propto f^{2d}$ vs $A_{\text{RF}} \propto (df)^2$). While both frequency dependant factors are quadratic, approximating the factors in (**??**), $f \propto A_{\text{aco}}$ is at least 4 orders of magnitude higher than $f \propto A_{\text{RF}}$

$$A_{\text{RF}}(d,f) \approx \left(\frac{4\pi df}{c}\right)^2 \text{ where } c \approx 3\times 10^8 \text{ms}^{-1} \tag{3.9}$$



(a) Linear Increasing

(b) Linear Decreasing

(c) Quadratic

(d) Isovelocity

Figure 3.3: Bellhop Model of Non-Linear Marine Shortest-Path Propagation in various Speed of Sound Profiles

$$10\log a(f) = 0.11 \cdot \frac{f^2}{1+f^2} + 44 \cdot \frac{f^2}{4100+f^2} + 2.75 \times 10^{-4} f^2 + 0.003 \tag{3.6}$$

Figure 3.4: Thorp's Absorption Model [?]

$$\begin{aligned} 10\log a(f) = & 0.106 \frac{t_1 f^2}{t_1^2 + f^2} e^{\frac{H^+ - 8}{0.56}} \\ & + 0.52 \left(1 + \frac{T}{43}\right) \left(\frac{S}{35}\right) \frac{t_2 f^2}{t_2^2 + f^2} e^{\frac{-D}{6}} \\ & + 4.9 \times 10^{-4} f^2 e^{-\left(\frac{T}{27} + \frac{D}{17}\right)} \end{aligned} \tag{3.7}$$

Where

$$t_1 = 0.78 \sqrt{\frac{S}{35} e^{\frac{T}{26}}}$$

$$t_2 = 42 e^{\frac{T}{17}}$$

Figure 3.5: Ainslie & McColm Absorption Model

$$10\log a(f) = A_1 P_1 \frac{t_1 f^2}{t_1^2 + f^2} + A_2 P_2 \frac{t_2 f^2}{t_2^2 + f^2} + A_3 P_3 f^2 \tag{3.8}$$

Where

$$A_1 = 1.03 \times 10^{-8} + 2.36 \times 10^{-10} \cdot T - 5.22 \times 10^{-12} \cdot T^2$$
$$A_2 = 5.62 \times 10^{-8} + 7.52 \times 10^{-10} \cdot T$$
$$A_3 = \left(55.9 - 2.39 \cdot T + 4.77 \times 10^{-2} \cdot T^2 - 3.48 \times 10^{-4} \cdot T^3\right) \times 10^{-15}$$
$$t_1 = 1.32 \times 10^3 (T + 273.1) e^{\frac{-1700}{T+273.1}}$$
$$t_2 = 1.55 \times 10^7 (T + 273.1) e^{\frac{-3052}{T+273.1}}$$
$$P_1 = 1$$
$$P_2 = 10.3 \times 10^{-4} \cdot P + 3.7 \times 10^{-7} \cdot P^2$$
$$P_3 = 3.84 \times 10^{-4} \cdot P + 7.57 \times 10^{-8} \cdot P^2$$

Figure 3.6: Fisher-Simmons Absorption Model

3.1.5 Ambient Noise Model

Ambient ocean noise can be assumed to be Gaussian with a continuous power spectral density in dB re µPaHz^{-1}, driven by four major factors, shown in **??**, where s is a shipping activity factor bounded from [0,1] and w is the surface wind speed in ms^{-1} [**?**].

Table 3.2: Contributing factors to Ocean Ambient Acoustic Noise

Source	Approximation
Turbulence	$10\log N_t(f) = 17 - 30\log f$
Shipping	$10\log N_s(f) = 40 + 20(s - 0.5) + 26\log f - 60\log(f + 0.03)$
Wind Driven Waves	$10\log N_w(f) = 50 + 7.5w^{\frac{1}{2}} + 20\log f - 40\log(f + 0.4)$
Thermal Noise	$10\log N_{th}(f) = 15 + 20logf$

3.1.6 Multipath effects

Refractive lensing and the multi-path nature of the medium result in line of sight propagation being extremely unreliable for estimating distances to targets (See **??** and **??**). The first arriving acoustic signal has as the very least curved in the medium, and commonly has reflected off the surface/seabed before arriving at a receiver, creating secondary paths that are sometimes many times longer than the first arrival path, generating symbol spreading over orders of seconds depending on the ranges and depths involved. Thus, the multi-path channel transfer function can be described by :

$$H(d,f) = \sum_{p=0}^{P-1} h(p) = \sum_{p=0}^{P-1} \Gamma_p / \sqrt{A(d_p,f)} e^{-j2\pi f \tau_p} \quad (3.10)$$

$$\text{where } \tau_p = d_p/c, c \approx 1500 \text{ms}^{-1}$$

where $d = d_0$ is the minimal path length between the transmitter and receiver, $d_p, p = \{1,...P-1\}$ are the secondary path lengths, Γ_p models additional losses incurred on each path such as reflection losses at the surface interface, and $\tau_p = d_p/c$ is the delay time.

3.1.7 Modelling and Simulation of the Acoustic Medium / Channel

Several toolkits exist in a variety of states that perform communications agent simulation, most notably the NS-2 / 3 family of frameworks and their add-ons. Some of these frameworks, such as SUNSET [**?**] and AquaTools [**?**], that are particularly proven in their capability in modelling static network performance, with less in built support for advanced, reactive node mobilities such as those involving collision or object avoidance.

Beyond the NS family, there are many other communications and simulation modelling systems such as OpNet++ [**?**] and MATLAB toolkits such as the AcTUP interface to the Ocean Acoustics Library, that primarily focus on simulation of the acoustic channel and contention issues without concentration on Underwater-specific Medium Access Control (MAC) protocols.

AUVNetSim is a simulation platform designed from the ground up with AUV operations in mind [**?**]. Including support for dynamic modular mobility and application behaviours, considering the stated context of an environmentally reactive UAN, AUVNetSim was tested and selected as a foundation upon which to build an exploratory network testing framework for this research.

In order to implement a collaborative, reactive, simulation suite, the SimPy [**?**] agent framework was used for "background" synchronisation.

As mentioned in the individual testing cases, transmission parameters for simulation were initially taken from and validated against results from **?**] and **?**].

3.2 Marine Operations, Payloads, Technologies, and Durations

The use and applications of AUVs has undergone a great expansion in recent years [**?**]. The primary application for AUVs has long been identified as the environmental monitoring of marine areas, and are actively being researched by a great range of industrial and defence sector applications, with secondary applications in the physical sciences and environmental research, which are summarised below[**?** **?**].

3.2.1 AUV operations and deployments

Hydrographic Survey

The use of AUVs in the place of manned-surface platforms or tethered undersea platforms enables greatly increased spatial and temporal sampling. Importantly, the separation of AUVs from the noisy sea surface enables much more efficient survey operations. This is particularly important when comparing to classical tow-line based measurements; where the mobility of the AUVs enables for much tighter-turning survey patterns or operation in inaccessible or hard-to-reach locations such as polar survey [**?**].

Another significant factor is cost; the daily cost of operating a manned vessel can be considerably higher than the costs of deploying, operating and recovering one or more AUVs with equivalent capabilities [**?**]. Additionally, the use of low-power "glider" AUVs has lowered the barrier to entry for extended mission types, such as persistent environmental survey, or open-ocean operations. Depth-hardened AUVs have also opened up the deepest parts of the oceans to exploration, with the onboard autonomy, imagery and Simultaneous Location and Mapping (SLAM) techniques allowing deep-dwelling survey AUVs to react to bottom-surface features without the need for a tight craft-to-surface control loop. The natural extension of these kind of applications is the use of AUVs on ice-covered planets such as Europa, where three-dimensional, autonomous navigation without an on-the-loop controller is vital for mission resource efficiency and success.

Hull and Infrastructure Inspection

Ongoing concerns regarding the security, safety and legality of international shipping has driven the application of AUVs to the area of near-surface hull and infrastructure inspections, looking for damage as well as devices such as limpet mines and other contraband. This use case puts a range of unique pressures on the AUV system; requiring highly accurate three-dimensional localisation and path-planning to clearly image the contours of a hull [**?**]. Similarly, with the increasing use and criticality of intercontinental undersea optical fibre connections, using AUVs for both the laying of and inspection of these cables is an exciting area of work [**? ?**].

Marine Petrochemical/Mineralogy

Oil and Gas industry requirements for high quality, low altitude bathymetry of seabed structures for infrastructure development (pipelines/drill platforms etc.) as well as monitoring of those structures over time (inspection etc.) is another significant application area, and a major driver of research investment. As in Hydrography, the mobility of AUVs is the biggest single advantage over classical platforms [**?**]. Additionally, recent advances in Synthetic Aperture Sonar (SAS) have provided invaluable sub-surface profile data over much wider areas for multi-spectral mineralogical analysis than previous sonar profilers [**?**].

Military

MCM Operations benefit greatly from, and significantly drive, AUV development; the ability to rapidly explore and covertly survey a potentially dangerous area without risking a human

operator is a major benefit both in Dedicated MCM (e.g. Large Area Hunting/ EOD Clearance) and in Organic MCM for Expeditionary Forces. This benefit applies to protection as well as incursion; the ability to rapidly deploy persistent survey of a valuable area such as a forward-operating harbour is increasingly essential, and as AUV technology, autonomy and security practices develop, this use is increasing. This Port Protection capability is particularly complex; teams of AUVs are expected to repeatedly survey an area and remain densely-connected enough to maintain end-to-end communications with all other nodes, in the face of an environment that is possibly not well surveyed initially, and includes dynamically moving obstacles (i.e. ships). In the remainder of this work this Port Protection scenario is used as a baseline for our simplified simulation context. Additional defence application areas include Anti-Submarine Warfare (ASW), Rapid Environment Assessment (REA), Navigational Aid/Force projection,

3.2.2 Localisation Technologies

Given the subsurface nature of most AUV operations, terrestrial localisation techniques such as GPS are unavailable (below $\approx 20cm$ depth). However, a range of alternative techniques are used to maintain spacial awareness to a high degree of accuracy in the underwater environment.

Long baseline (LBL)

Long-baseline localisation systems use a series of static surface/cable networked acoustic transponders to provide coordinated beacons and (usually) GPS-backed relative location information to local subsurface users. Such systems can be accurate to less that 0.1m or better in ideal deployments and are regularly used in controlled autonomous survey environments such as harbour patrol operations where the deployment area is bounded. However, the initial set-up and deployment required in advance of any AUV operation makes LBL difficult to utilise in unbounded or contended areas. LBL systems can also be deployed on mobile surface platforms in the area (ships or buoys for example), but these applications put significant computational pressure on the end-point AUV and have greatly reduced accuracy compared to ideal deployments [**?**].

Doppler Velocity Log (DVL)

Doppler Velocity Logging involves the emission of directed acoustic "pings" that reflect off sea bed/surface interfaces that, when received back on the craft with multi-beam phased array acoustic transducers can measure both the absolute depth/altitude (z-axis) of the craft and through directional Doppler shifting, the relative (xy-translative) motion of the craft since the ping. While classical DVL was highly sensitive to shifting currents in the water column, advances in the development of Acoustic Doppler Current Profiling has turned that situation on it's head, enabling the compensation-for and measurement-of water currents down to the sub-meter level [**?**].

Inertial Navigation System (INS)

Inertial navigation systems use gyroscopic procession to observe the relative acceleration of a mobile platform. This reference-relative monitoring is particularly useful in the underwater environment, as it detects the motion of AUVs as they are carried by the water itself. Bias Drift is a significant problem for INS systems operating over longer (hundreds of metres) distances, as they usually have some minimal amount of directional bias, that incurs a cumulative effect over time without assistance. Several sensor synthesis processes have been demonstrated which combine information from INS along with DVL data to improve localisation into the sub-decimeter level [**? ? ?**].

Simultaneous Location and Mapping (SLAM)

Simultaneous Location and Mapping is the process of iteratively developing a feature-based model of an environment, and to use the relative movement within that modelled environment to obtain estimates of absolute positioning. SLAM has been most well developed in the contexts of either visual-based inspection using cameras, or LIDAR-style distance triangulation, however the same principles have been successfully applied using marine sonar readings, providing sub-meter accuracy, real-time, feature-relative localisation information that is (for the most part) environmentally agnostic [**?**].

In summary, current technology reliably enables AUVs to localise to a sub-metre accuracy in most areas of application.

3.2.3 Example Maritime Autonomous Systems, Platforms and Operational Limitations

Kongsburg REMUS/HUGIN ranges

The REMUS range of AUV platforms have been very popular in research and UAN application prototypes due to their relatively small size and high level of reconfigurability. The basic configurations of the REMUS 100 configuration consist of a single pressure vessel, 0.2m in diameter and 1.6m long, weighting in at $37kg$, rated to operational depths of 150m. This package includes DVL, Conductivity, Temperature and Depth of ocean (CTD), Underwater Videography, LBL and onboard computing power suitable for low LOA independence, with onboard Li-on battery packs rated to provide up to 10-hours of cruising operational endurance. These capabilities can be extended through the addition of further modular extensions through the REMUS range, such as the REMUS 600, rated for up to 600m depth and a cruising endurance of 45 hours, or the REMUS 6000, rated for up to 6km with 22 hours duration.

The HUGIN 1000 is a high-resolution extension to the REMUS range, characterised by it's default payload of a High Definition SAS, co-designed with the Norwegian Defence Research Establishment (FFI) for MCM operations, with a dynamic depth rating up to 3km and 24 hour cruising endurance (17 hours with continuous SAS engagement)

The Konsburg range also include range specific Launch And Recovery System (LARS)

Figure 3.7: HUGIN AUV mounted on LARS

NOC Autosub

Developed under the UK's National Oceanography Centres Marine Autonomous and Robotic Systems group, the Autosub family of AUVs is similar in many ways to the REMUS deployment profile; with long range and deep-ocean variants, operating at depths up to 600 km for up to 36 hours (however, this configuration leaves it with a cruising speed of $0.4\,\mathrm{ms}^{-1}$)

Figure 3.8: NOC Autosub 3 being deployed off the Pine Island Glacier

University of Washington SeaGlider

Taking a fundamentally different approach to underwater mobility for targeting depth-variant environmental studies, the SeaGlider eschews classical propulsion to use it's downward-facing structure and fins to use it's weight/buoyancy to propel itself. At 1.8 m long and weighting only 52 kg, this highly portable AUV can cover ranges up to 4600 km with 650 1 km dive segments at a rate of $0.25\,\mathrm{ms}^{-1}$.

USN Sea Hunter

The Sea Hunter is an autonomous unmanned surface vehicle (USV) launched in 2016 and is undergoing seatrials as part of the DARPA Anti-Submarine Warfare Continuous Trail Unmanned Vessel program, with a top speed of $50\,\mathrm{kmh}^{-1}$, weighing 122 t. While unarmed during its sea trials, the *Sea Hunter* will be armed and used for ASW and MCM duties, operating at a tiny fraction of the standard operating costs of a littoral destroyer.

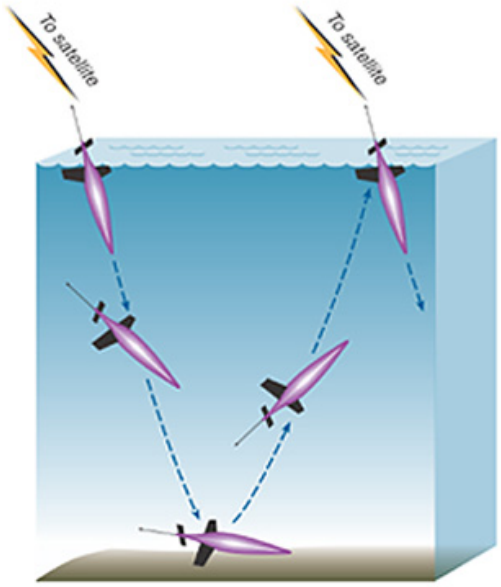


Figure 3.9: "Flight path" of the UW SeaGlider

Figure 3.10: Initially manned deployment of the unmanned *Sea Hunter*. U.S. Navy photo by John F. Williams/Released

3.2.4 Need for Trust in Maritime Networks

Given the breadth of the threat space in MANETs, many strategies for mitigating these risks have been proposed in a matrix-basis; i.e. you can't cover all eventualities with one tool. **??** summaries some of these general strategies or "solutions" for range of threats identified specifically in the context of the deployment of MANETs in a tactical/defence context, originally based on **?**] but with some contextual alterations with modifications discussed below.

In this context of this table, Vulnerability is an assessment of how available a particular threat is to a generic attacker, Impact is an assessment of the in-system affects of a successful attack, and Risk is a qualitative assessment of the likelihood of exploitation of an attack. **?**] originally lowered the assessment of vulnerability of Eavesdropping, citing the tactical use of non-standard MAC as a barrier to attack, however considering the increasing commoditisation of tactical hardware across the world, and increasing application of Consumer Off-The-Shelf (COTS) hardware, this is no longer a fair assessment. Similarly, the assessment that the Risk of Data Corruption is low is arguable on the basis of simplistic spread spectrum jamming but this assessment is unchanged.

Table 3.3: Risks and Threat Mitigation Strategies for MANETs, extended from [?]

Threat	Vulnerability	Impact	Risk	Mitigation
Resource Depletion / DoS	Low-High	High	Low-High	Layer Specific Mechanisms [? ?]
Eavesdropping	Medium	High	Low	Cryptography [?]
Masquerade	Low	Very High	Medium	Trust Systems and Cryptography [?]
Data Corruption	Low	High	Low	Cryptography
Traffic Analysis	High	Low	Medium	Obfuscation [?]

Resource Depletion, or DoS has a very wide ranging definition, from network-level attacks to saturate a communications channel or the computing resources of routing nodes, or the exploitation of a power-control loop to induce a node to waste energy with overly-high-powered communications, to the intentional geographic misleading of nodes to induce a similar power-drain on locomotive systems, through tactics such as location spoofing or GPS denial [?].

From ??, it is assessed that the highest overall threats are those of Resource Depletion and Masquerading. Within this threat context, the general optimisation of any TMF would be to prefer high but fair overall network throughput, while minimising delay, PLR, and power usage.

3.3 Conclusion

As Autonomous Underwater Vehicle (AUV) platforms become more capable and economical, they are being used in many applications requiring trust. These applications are using the collective behaviour of teams or fleets of these AUVs to accomplish tasks [?]. With this use being increasingly isolated from stable communications networks, the establishment of trust between nodes is essential for the reliability and stability of such teams. In the next chapter, the use of Trust methods developed in the terrestrial MANET space will be re-appraised for application within the challenging underwater communications channel.

Chapter 4

Assessment of TMF Performance in Marine Environments

4.1 UANs as MANET analogue

As MANETs grow beyond the terrestrial arena, their operation and the protocols designed around them must be reviewed to assess their suitability to different communications environments to ensure their continued security, reliability, and performance. With demand for smaller, more decentralised MANET systems in a range of domains and applications, as well as a drive towards lower per-unit cost in all areas, TMFs are increasingly applied to resource constrained applications, as the benefits and efficiencies these systems present are significant. Many UANs use MANET architectures, however the marine environment presents new challenges for trust management frameworks that have been developed for use in conventional (i.e. Terrestrial RF) MANETs. These increasingly decentralised applications present unique threats against trust management [**?**].

Previous research has established the advantages of implementing TMFs in 802.11 based MANETs, particularly in terms of preventing selfish operation in collaborative systems [**?**], and maintaining throughput in the presence of malicious actors [**?**]

To date this work has been limited to terrestrial, RF based networks, which, as discussed in **??**, is a much more favourable communications environment compared to the marine environment of UANs, where extreme communications challenges are present (propagation delays, frequency dependent attenuation, fast and slow fading, refractive multipath distortion, etc.). As a result of these challenges, in underwater environments, communications is both sparse and noisy. That is to say that long delays create high susceptibility to contention blocking, requiring relatively low channel occupancy and significant back-offs and significant retransmission penalties in the face of a multitude of noise sources over inconsistent, non-linear, multi path transmissions. Therefore the observations about the communications processes that are used to generate the trust metrics, occur much less frequently, with much greater error (noise) and delay than is experienced in terrestrial RF MANETs. Beyond the constraints of the communications environment, knock on pressures in battery capacity, on-board processing, and locomotion simultaneously present

opportunities and incentives for malicious or selfish actors to appear to cooperate while not reciprocating, in order to conserve power for instance. These multiple aspects of potential incentives, trust, and fairness do not directly fall under the scope of single metric trusts discussed previously in **??**, and this context indicates that a multi-metric approach may be more appropriate.

As such, the use of trust methods developed in the terrestrial MANET space must be re-appraised for application within the underwater context [**?**].

This chapter is primarily concerned with the analytical establishment of hard trust within a topologically dynamic network of mobile autonomous actors. It will be shown that single metric trust systems are not directly suitable for the marine context in terms of the different threat and cost scenario in that environment. These single metric TMFs provide malicious actors with a significant advantage if their activity does not impact that metric.

For the purposes of this work, from those TMFs discussed in **??**, Hermes trust establishment, OTMF and MTFM are selected as indicative single and multi metrics frameworks for comparison, as Hermes captures the core operation of a pure single metric assessment methodology and OTMF provides a comparison that combines assessments from across nodes to develop trust opinions. MTFM is also included as an example of an existing multi-metric TMF that looking purely at the communications domain.

4.2 Modelling of UAN network

4.2.1 Mobility, Topology, and Communications

Four mobility patterns are initially investigated:

1. All Nodes Static
2. Malicious node mobile
3. Malicious node mobile, all other nodes static
4. All nodes mobile

For this case, the mobility model used is a random walk on the nodes modelled kinematic response, i.e. the node periodically picks a spherically normalised random direction in the XY plane. Maximum node speed (limited by kinematic acceleration/turning constraints) is $1.5ms^{-1}$[**?**].

The six nodes are initially arranged as per Fig. **??** with each node on average 100m from each other as per **?**]. The use of six nodes and the particular layout enables the investigation of the three trust relationships based on minimum path topologies, such that the node generating the trust assessments, n_0 has Direct, Recommendation, and Indirect trust assessments of n_1 available to it from itself, [n_2,n_3], and [n_4,n_5] respectively. (See Section **??**)

Collaborations with NATO Centre for Maritime Research and Experimentation (CMRE) in La Spezia, and Defence Science and Technology Laboratorys (DSTLs) Naval Systems Group inform that this is a practical team-size for environmental and defence applications.

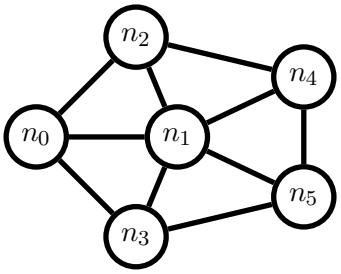


Figure 4.1: Initial layout with nodes spaced an average of 100 m apart

Simulation Background

Simulations were conducted using a Python based simulation framework, SimPy [**?**], with a network stack built upon AUVNetSim [**?**], with transmission parameters (**??**) taken from and validated against existing studies [**? ? ?**]

Given the differences in delay and propagation between RF and marine networks, it would not be expected that the same application rates (e.g. packet emission rates or throughput) and node separations are equally stable in this environment. Therefore, a zone of performance is characterised within which the network has stable operation.

Table 4.1: Comparison of system model constraints as applied between Terrestrial and Marine communications

Parameter	Terrestrial	Marine
Simulated Duration	300 s	18000 s
Trust Sampling Period	1 s	600 s
Simulated Area	0.7 km^2	0.7 to 4 km^2
Transmission Range	0.25 km	1.5 km
Physical Layer	RF(802.11)	Acoustic
Propagation Speed	3×10^8 ms^{-1}	1490 ms^{-1}
Center Frequency	2.6×10^9 Hz	2×10^4 Hz
Bandwidth	22×10^6 Hz	1×10^4 Hz
MAC Type	CSMA/DCF	CSMA/CA
Routing Protocol	DSDV	FBR
Max Speed	5 ms^{-1}	1.5 ms^{-1}
Max Data Rate	5×10^6 $bits^{-1}$	$\approx$240$bits^{-1}$
Packet Size	4096 bit	9600 bit
Single Transmission Duration	10 s	32 s
Single Transmission Size	1×10^7 bit	9600 bit

4.2.2 Establishing Scale Factors in Communications Rate

In this section the simulated communications environment is characterised to establish an optimal packet emission rate for comparison against [?]. This optimal emission rate is taken to be an emission rate that provides reasonable network stability and protection from network saturation. Network saturation is the point at which a network can no longer successfully deliver the offered load[1] presented to it to the relevant destinations (throughput), and is characterised by a peak and a subsequent decline in the throughput of the network when varying the packet emission rate.

Formally, this saturation rate occurs if

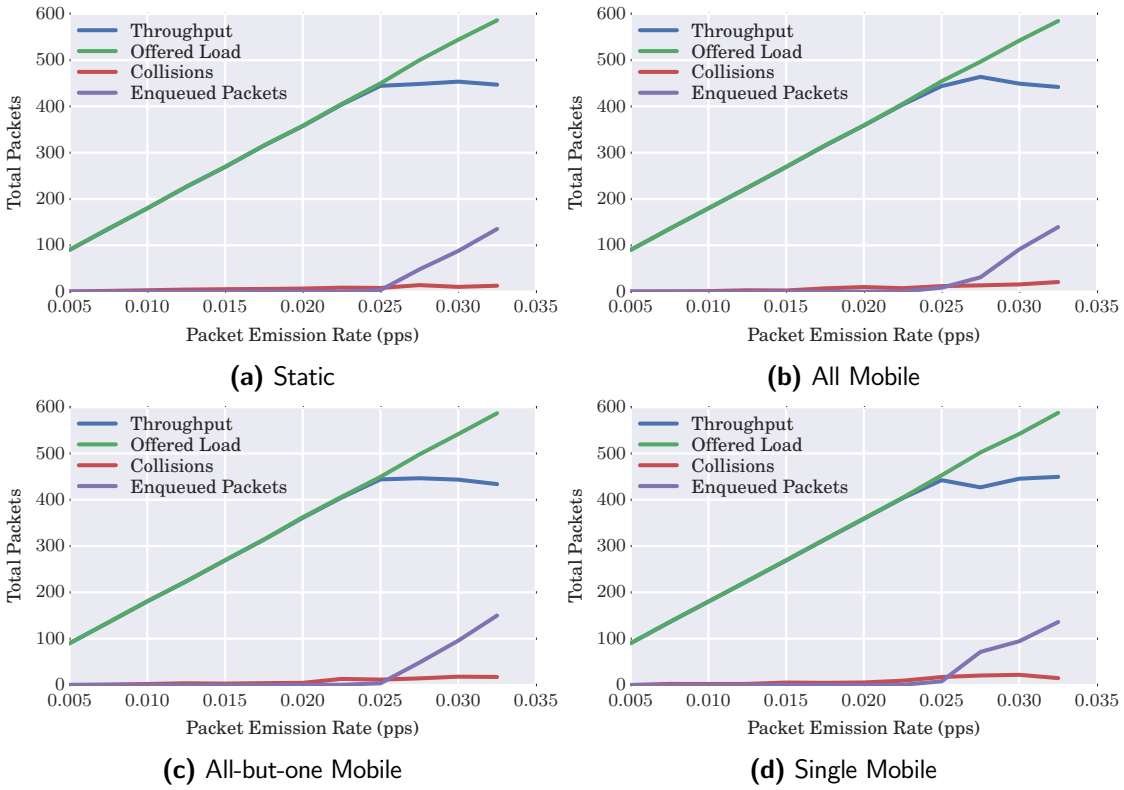
$$N\lambda_s > \mu_{\max} \tag{4.1}$$

In order to establish the point at which the network becomes saturated due, a range of packet emission rates were explored between 0.01 packets per second (pps), equivalent to 96 bits of offered load per node, up to $0.07pps$ ($672bps$ per node). Initial node separation was set as per ?] at 100m, and each simulation is run 16 times, with each instance modelling a 8 hour mission time. This configuration and duration are specified to correlate to previously discussed mobile collaborative port protection scenarios from ??.

Looking first at the Static mobility case, where all nodes are stationary; from ?? it is already clear that the throughput curve, exhibits a saturation point close to 0.025 pps. Similarly in ??, the precipitous drop in packet delivery probability beyond 0.025 pps, indicating that this is a strong candidate value for an upper-limit to the safe operating zone in terms of packet emission in the small static case. From ??, raising packet emissions above $0.025pps$ results in a significant increase in end-to-end delay. As per ??, the Carrier Sense Multiple Access (CSMA) based MAC incurs a certain amount of control overhead in the form of Request To Send (RTS) packets, when a node attempts to acquire time in its neighbourhood. In ??, the ratio of Control/Data packets increases linearly up to 1.5 until just before $0.025pps$, and then accelerates to almost 2.5, further demonstrating that the network has become critically congested. It is worthwhile noting that in ?? that even as the saturation point is passed, packet collisions do not significantly increase, and that the saturation is in fact driven by contention in the medium rather than congestion-collisions.

Results are also included from the remaining mobility cases (all nodes mobile; all-but-one node mobile; single mobile node), however from Figs. ??- ??, the throughput threshold behaviour is qualitatively similar regardless of mobility for this initial node separation.

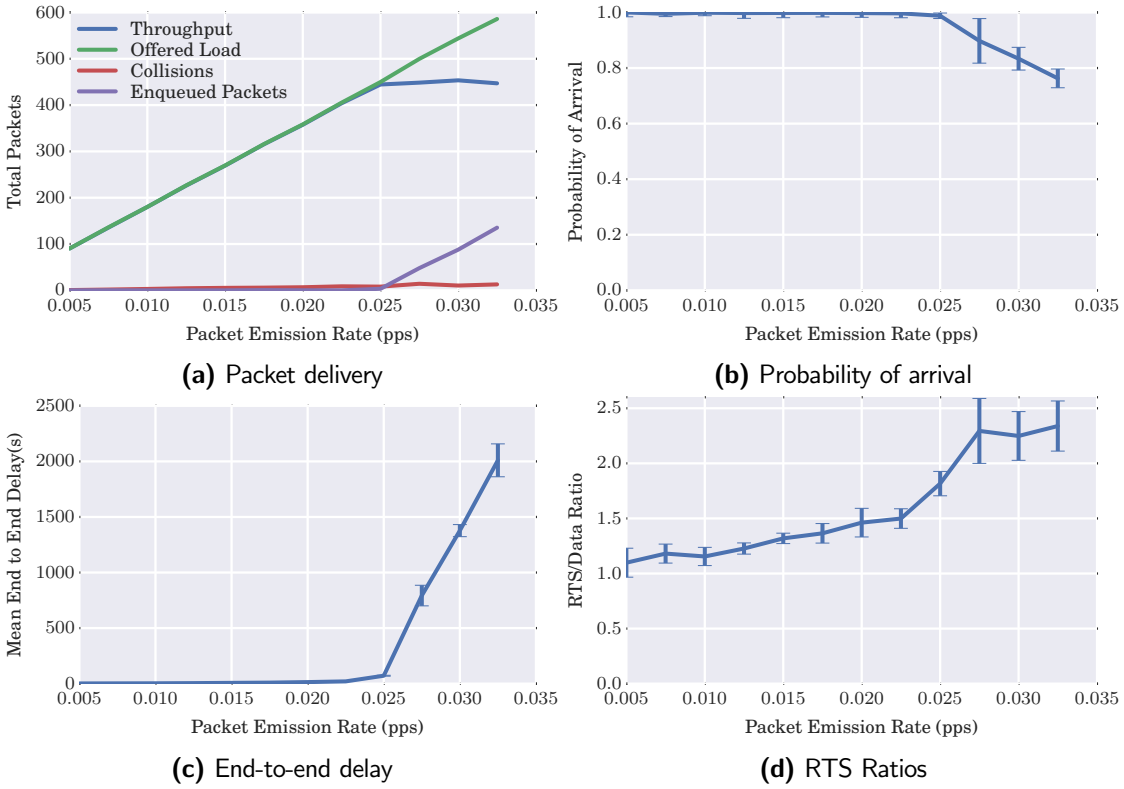
[1] It will become important to note that Offered Load in this case includes packet retransmissions



(a) Static **(b)** All Mobile

(c) All-but-one Mobile **(d)** Single Mobile

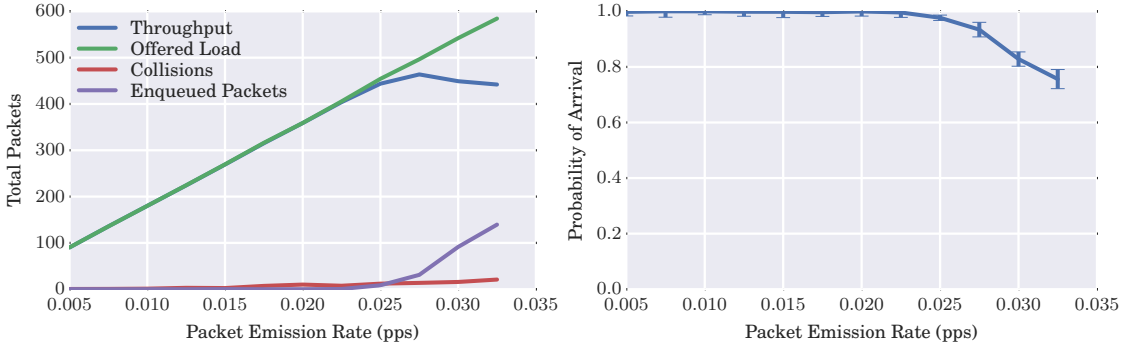
Figure 4.2: Throughput performance overview for all mobilities under varying emission rates



(a) Packet delivery **(b)** Probability of arrival

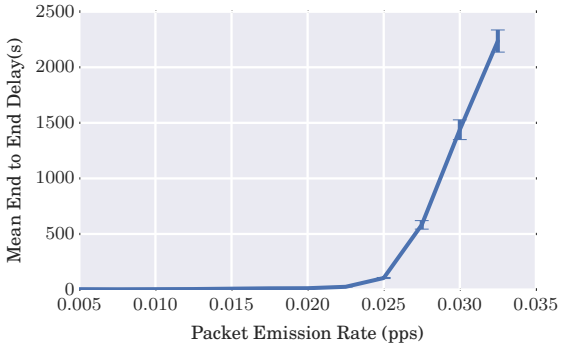
(c) End-to-end delay **(d)** RTS Ratios

Figure 4.3: Network performance varying packet emission rates for the static case

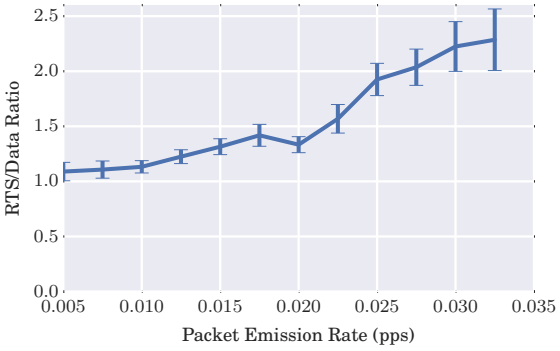


(a) Packet delivery

(b) Probability of arrival

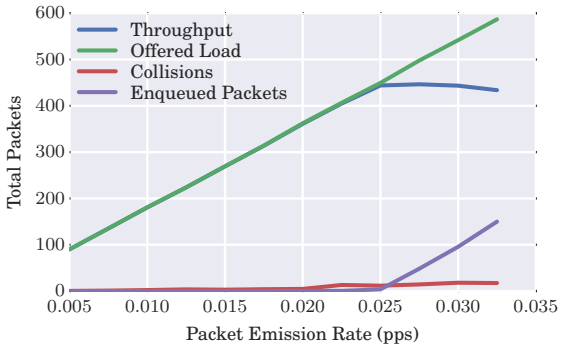


(c) End-to-end delay

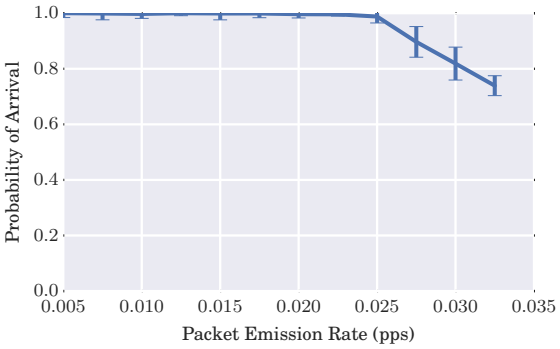


(d) RTS Ratios

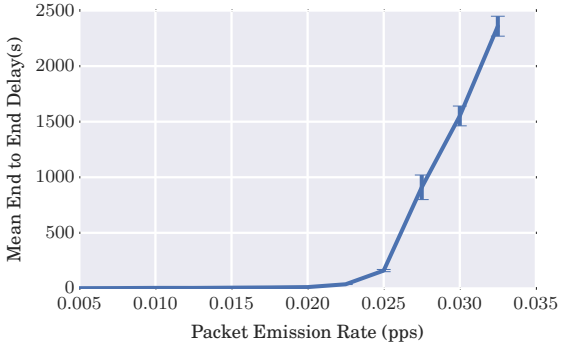
Figure 4.4: Network performance varying packet emission rates for the all mobile case



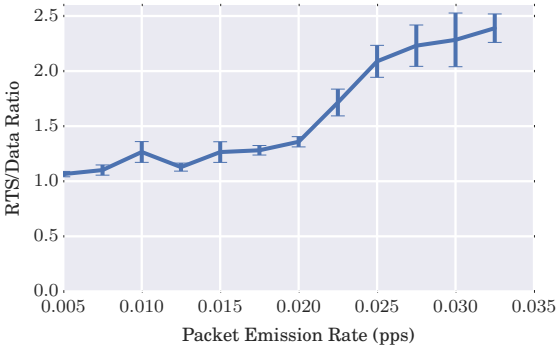
(a) Packet delivery



(b) Probability of arrival

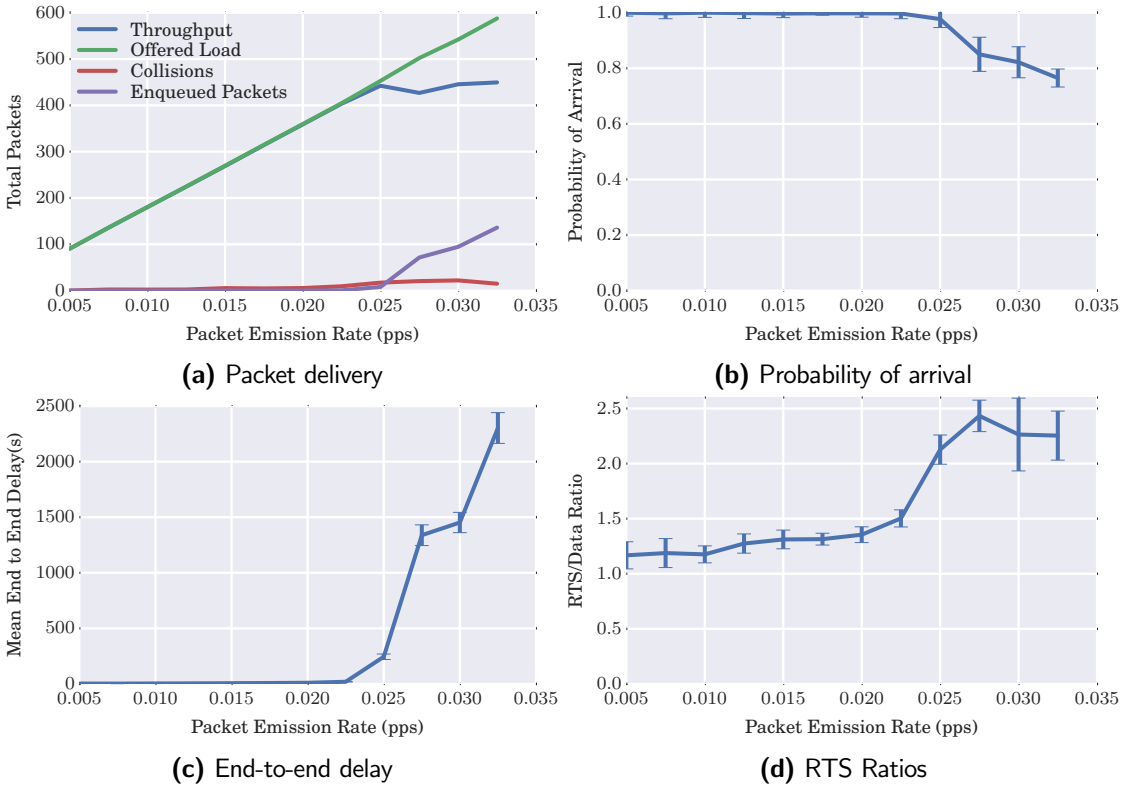


(c) End-to-end delay



(d) RTS Ratios

Figure 4.5: Network performance varying packet emission rates for the all-but-one mobile case



(a) Packet delivery

(b) Probability of arrival

(c) End-to-end delay

(d) RTS Ratios

Figure 4.6: Network performance varying packet emission rates for the single mobile case

4.2.3 Scale Factors in Physical Node Distribution

In this section the effect of node-separation scaling on communications operation is characterised for comparison against [?]. This is particularly important considering the significant scale factor differences in terms of the speed of propagation in the medium, and the range of potential desired operation.

From ??, the operating transmission range of acoustic is ≈ 6 times further than 802.11, indicating that a suitable operating environment will have an area $\approx \sqrt{6}$ times the area of the 802.11 case. Therefore, a reasonable experimental range would have an upper bound of performance around this scaling factor, where nodes are approximately $400m$ apart.

According to ?], RTS/Clear To Send (CTS) handshake functionality cannot operate well as interference protection at node separations beyond 0.56 times the transmission range[?]. In the case of marine acoustic transmission at the stated power output, above $1500\text{m} \times 0.56 = 840\text{m}$, handshake overheads should begin to dominate channel access. This is due to reduced channel availability due to collisions, which are then due to a much longer potential contention period between nodes.

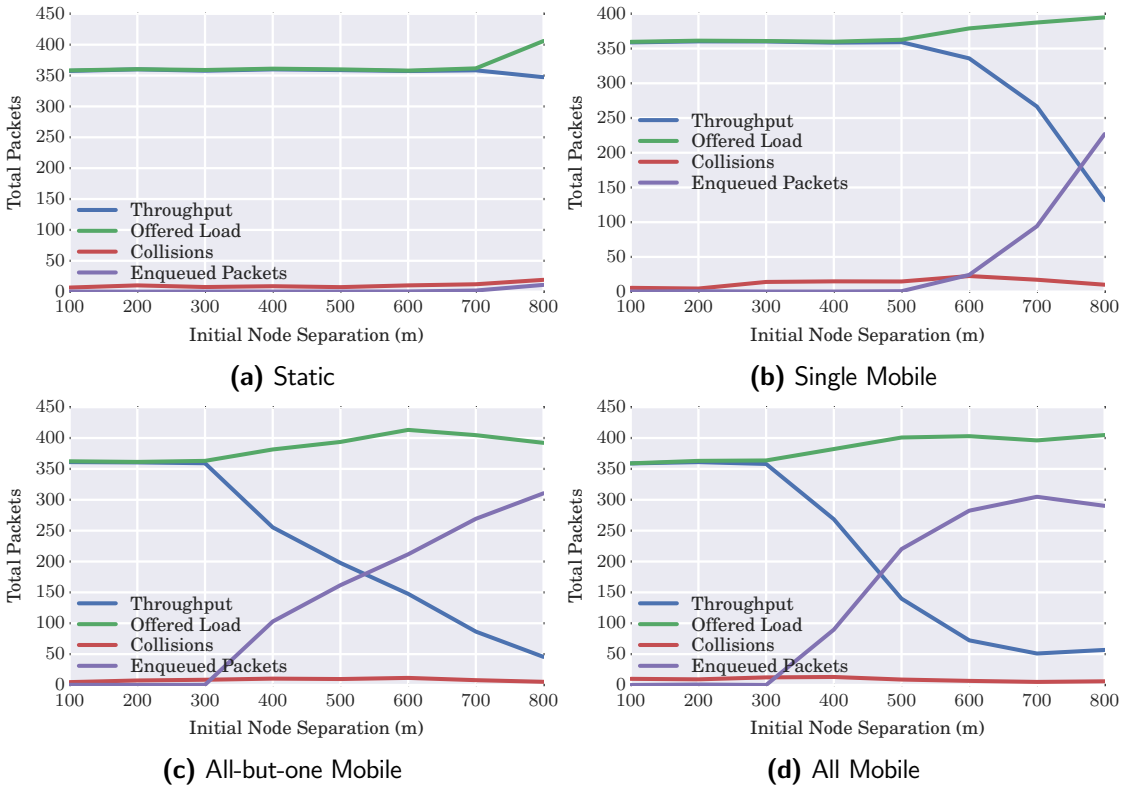
A reasonable range around this is to scale from $100m$ apart on average to 800m, and from the previous section, a packet emission rate of $0.02pps$ (slightly below the $0.025pps$ saturation threshold) is used to explore this space. The "environment" of the simulations is also scaled in accordance with the node scaling, based on an initial environmental "water-box" of 1km for the 100m node separation, i.e. the water-box is consistently ten times larger than the initial node separation.

In the case where all nodes remain static, increasing node separation does not significantly impact throughput, delay, delivery probability or RTS ratios until rising above $700m$ (Fig. ??), nearly double the initial estimate of where an appropriate separation zone would be.

The other mobility cases tell a very different story; as can be seen in ??, where adding a single mobile node to the network induces a saturation-style response at $500m$, and this drops further in ?? and ??, reducing the separation of saturation at this emission rate to just $300m$.

Another aspect of these results to highlight is that the Offered Load presented to the network *increases* beyond the collapse of the throughput curve. This indicates that there is a subtly different saturation behaviour with respect to separation than the simple congestion argument with respect to packet emission rate; packets are simply taking too long to cross the increasingly sparse network and in-transit packet routes are logically disconnected and require retransmission.

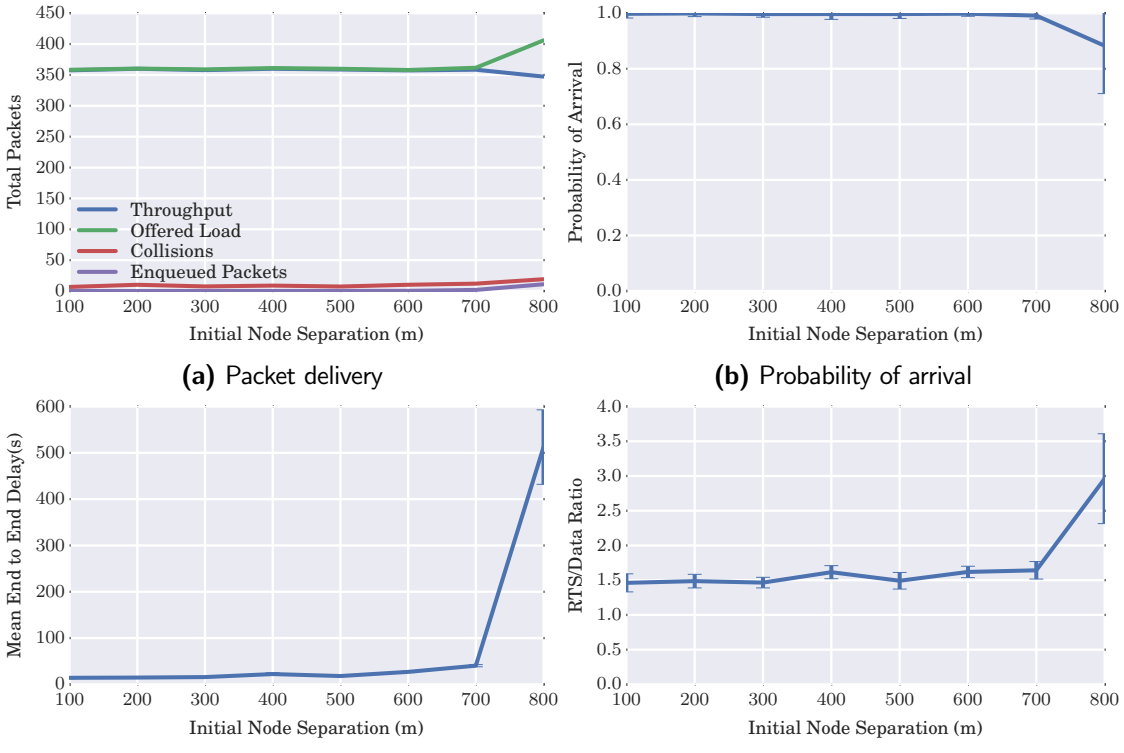
4.2.4 Combined Scale Factor Analysis



(a) Static **(b)** Single Mobile

(c) All-but-one Mobile **(d)** All Mobile

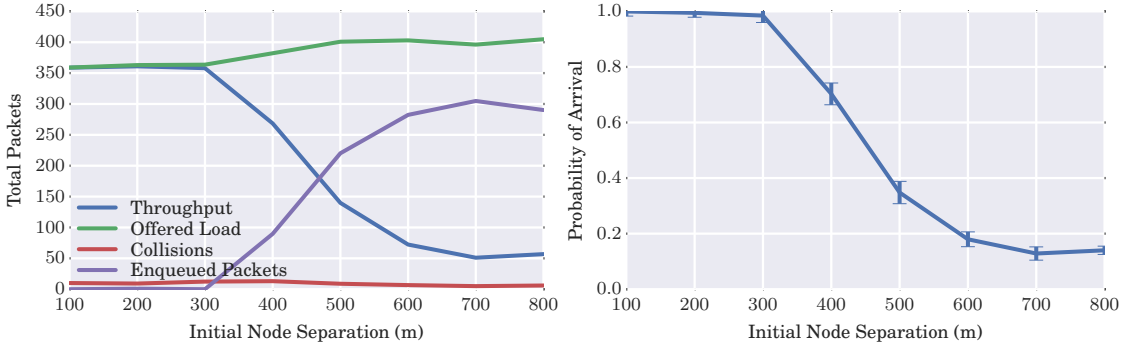
Figure 4.7: Throughput performance overview for all mobilities under varying separation



(a) Packet delivery **(b)** Probability of arrival

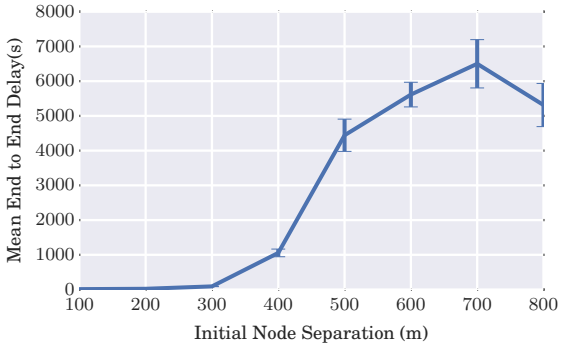
(c) End-to-end delay **(d)** RTS Ratios

Figure 4.8: Network performance varying node separation for the static case

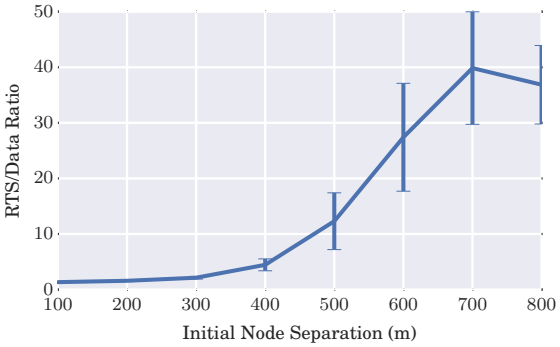


(a) Packet delivery

(b) Probability of arrival

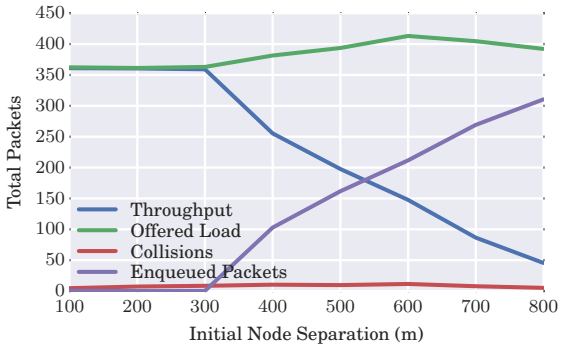


(c) End-to-end delay

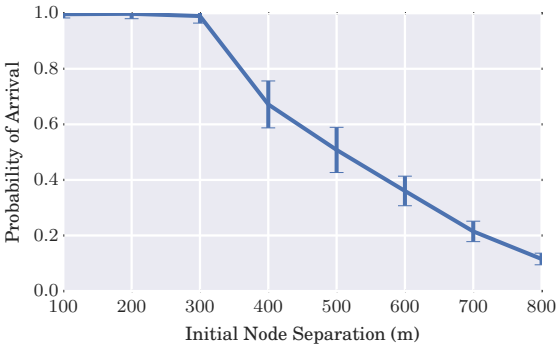


(d) RTS Ratios

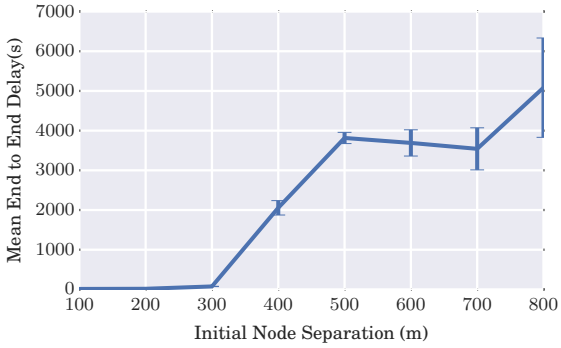
Figure 4.9: Network performance varying node separation for the all mobile case



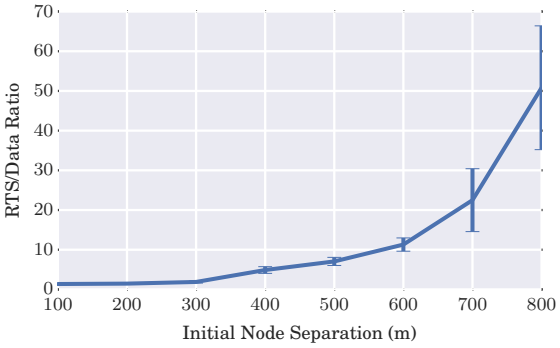
(a) Packet delivery



(b) Probability of arrival



(c) End-to-end delay

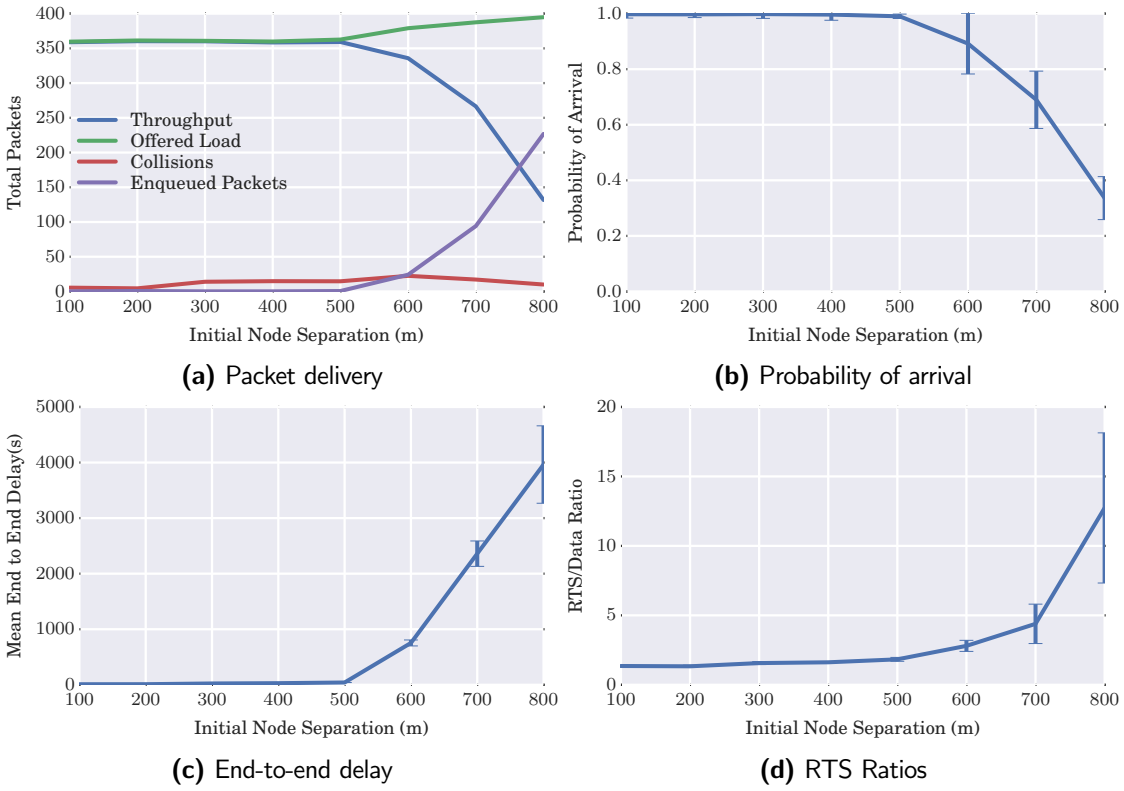


(d) RTS Ratios

Figure 4.10: Network performance varying node separation for the all-but-one mobile case

Table 4.2: Tabular view of data from ??, including ideal propagation time

Initial Node Separation (m)	Delay(s)	Probability of Arrival	RTS/Data Ratio	Ideal Delivery Time(s)
100	10.3551	0.9977	1.3546	1.0314
200	11.1631	0.9973	1.3322	1.1029
300	24.2225	0.9983	1.5650	1.1743
400	29.4864	0.9965	1.6210	1.2457
500	41.7093	0.9904	1.8331	1.3171
600	753.4040	0.8922	2.8038	1.3886
700	2360.0826	0.6899	4.3889	1.4600
800	3963.9830	0.3360	12.7323	1.5314



(a) Packet delivery **(b)** Probability of arrival

(c) End-to-end delay **(d)** RTS Ratios

Figure 4.11: Network performance varying node separation for the single mobile case

Its clear from the previous results that the relationship between emission rates, separations and mobilities is tightly coupled and not totally clear cut. To arrive at a more optimal operating region, a coupled analysis is performed across both emission rate and initial separation distance.

Given what has been discussed so far; it's clear that in identifying an appropriate operating region, it is important to not only ensure throughput, but that that throughput is timely. For instance, in ?? (tabulated in ??), a small increase in separation beyond the apparent throughput-peak at $500m$ to $600m$, which constitutes an increased ideal marine acoustic "time of flight" between nodes by $0.02s$, increases the average actual delay by 1800%.

To capture these performance requirements, the feature scaled product of Throughput and Delay is taken and plotted against rate and separation in ??.

$$V = |S] \times (1 - |D|) \tag{4.2}$$

For each scenario, the observed Throughput across the network (S in bytes) is normalised across all observations (i.e. each combination of Node Separation and Emission Rate), as is average end-to-end Delay (D). The normalised delay is inverted ($1-|D|$) and the product of this and the normalised throughput is used as the basis of a two-dimensional linear interpolation shown in ??.

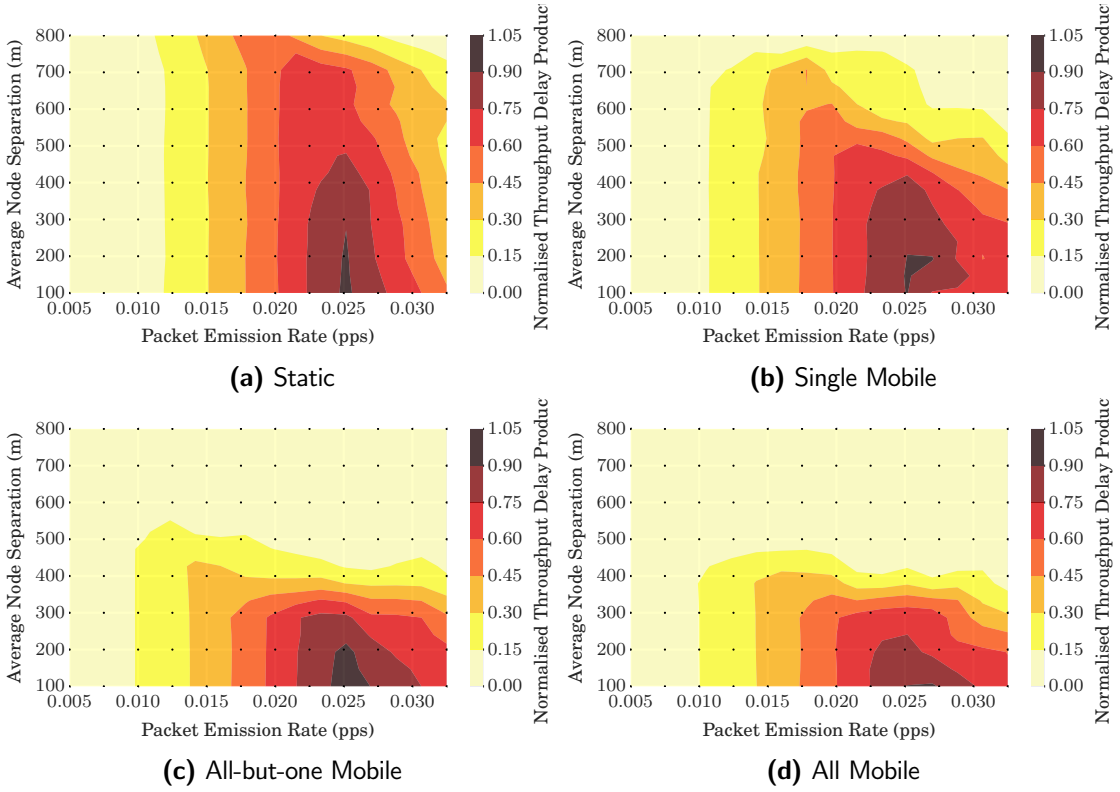


Figure 4.12: Normalised Throughput-Delay Product for all mobilities under varying separation and emission rate

4.2.5 Summary

An appropriate safe operating zone for marine communications has been established by investigating the impact of variations of the communications rate and physical distribution across the mobility scenarios.

These findings can be summaries as that when the separation is increased, the emission rate at which the network becomes saturated decreases, reducing overall throughput. This throughput degradation is tightly coupled with the mobility, as increasing mobility leads to increasing delays as routes are constantly broken, re-advertised and re-established. For instance, where all nodes are static, significant drops in throughput are not seen until node separation approaches 800m, nearly double the initial estimate. However, when all nodes are randomly walking the saturation point collapses from 0.025*pps* at 300m to 0.015*pps* at 400m. These results indicate that a good area to continue operating in for a range of node separations is at 0.015*pps*, and that a reasonable position scaling is from 100m to 300m, beyond which communication becomes increasingly unstable, especially in terms of end-to-end delay. These results are similar to related simulation work [**?** **?** **?**], and is to be expected in such a sparse, noisy, and contentious environment. It should be noted that these rates are not what would be expected in a single-node/single-operator environment that most current AUV operations inhabit, as without any channel contention and mixed-delay effects, data-rates many times higher than this would be expected. However, few practical experiments have been performed that are suitable for direct comparison, as this context relies on multiple available nodes in a dynamic MANET topology with differential mobility.

The results from **??** and **??** show that the single-node differential mobility models don't capture the reality of the network in the proposed port-protection context. The reason for this is that in these single-differential mobility combinations, the node targeted for misbehaviour (n_1) will already be behaving differently compared to the rest of the network regardless of the misbehaviour. A future extension to this work could be to look at differential ratios of static/mobile nodes in alternative scenarios, such as in data-muling applications or WSN.

4.3 Operation of TMFs in UANs

We are primarily concerned with the direct trust relationship between n_0 and n_1, i.e. n_0's assessment of the trustworthiness of n_1, or $T_{1,0}$.

?] introduced a range of misbehaviours, including modification of the packet loss rate of routing nodes and limiting throughput on a per-link basis as well as a selection of combined misbehaviours. Given that the established links are already heavily constrained, such attacks would severely impact the general performance of the network beyond the scope of simple selfishness. These direct malicious behaviours effectively trigger saturation collapses in operating regions of the network that should be stable.

Therefore, two more subtle misbehaviours to investigate are;

1. Malicious Power Control (MPC), where n_1 increases its transmit and forwarding power by 20% for all nodes *except* communications from n_0 in order to make n_0 appear to be selfishly conserving energy to the rest of the team, while n_1 itself appears to be performing very well.

2. Selfish Target Selection (STS), where n_1 preferentially communicates, forwards and advertises to nodes that are physically close to it in effort to reduce its own power consumption.

4.4 Simulation Results and Discussion

Having established a safe operating range for comparison at 300m average separation and an emission rate of $0.015pps$, each of the three selected behaviours (Fair, Malicious Power Control (MPC), Selfish Target Selection (STS)) are performed in both the static and mobile scenarios. We select a trust assessment period of 10 mins for a five hour mission to scale in comparison to relative bitrates experienced ($1Mbps$ vs $\approx 15bps$).

The six metrics used for grey assessment are; transmitted and received throughput and power, delay, and PLR as calculated by aborted and unacknowledged, transmissions. Compared to [**?**], this metric set lacks a data rate quantity as the network is not dynamically adjusting bandwidth. In context of GRC generation (**??**), the best sequence g was selected using the lowest PLR, delay, and powers, and the highest throughputs, and the worst sequence, b the inverse of these metrics, reflecting the observations made in **??**.

The particular factors under discussion are the relative performance of MTFM against OTMF and Beta with respect to statistical stability across mobilities and in responsiveness to changing network behaviour. We establish a similar result set by initially tracking the resultant trust values established by MTFM in the pair of mobility scenarios, shown in Fig. **??**. We are also concerned with the opinions of n_1 provided to n_0 by other nodes, where $[T_{2,1},T_{3,1}]$ and $[T_{4,1},T_{5,1}]$ denote the sets of recommendation and indirect trust assessment respectively.

We also include aggregate assessments; $T_{N,1}^{\text{Avg}}$, the unweighted mean of direct trust assessments of n_1 from all nodes and $T_{0,1}^{\text{MTFM}}$, the final MTFM trust assessment value based on both network topology with respect to n_0 and whitenization from (**??**).

The variability in assessment is coupled to mobility; in the static case (Fig. **??**), the nodes exhibit relatively consistent distributions. In the full mobility case, shown in Fig. **??**, this subjective variability is greatly increased. As the topology is highly dynamic, delays due to re-establishing routes can be very large, perturbing the trust value. The $T_{0,1}^{\text{MTFM}}$ displays a significantly reduced variation than those of the individual subjective observations in all cases, even when compared to the unweighted average, $T_{N,1}^{\text{Avg}}$. This demonstrates T_{MTFM}'s value as an aggregating trust assessment in such sparse and noisy environments. Further, in Fig. **??** a much higher variability in assessment is observed in $T_{0,1}$, correctly indicating that there is something wrong with the relationship between n_0 and n_1.

4.4.1 Comparison between MTFM, Hermes and OTMF

As per **?**], "fair" scenarios were also performed with no malicious behaviour, applying OTMF and Hermes assessment as well as MTFM, providing like-for-like comparison of assessment.

The use of FBR and a CSMA with Collision Avoidance (CSMA CA) MAC scheme from AUVNetSim [**?**] in our simulation mitigates a significant number of packet losses through collision avoidance and contention handling, leading to the situation that the

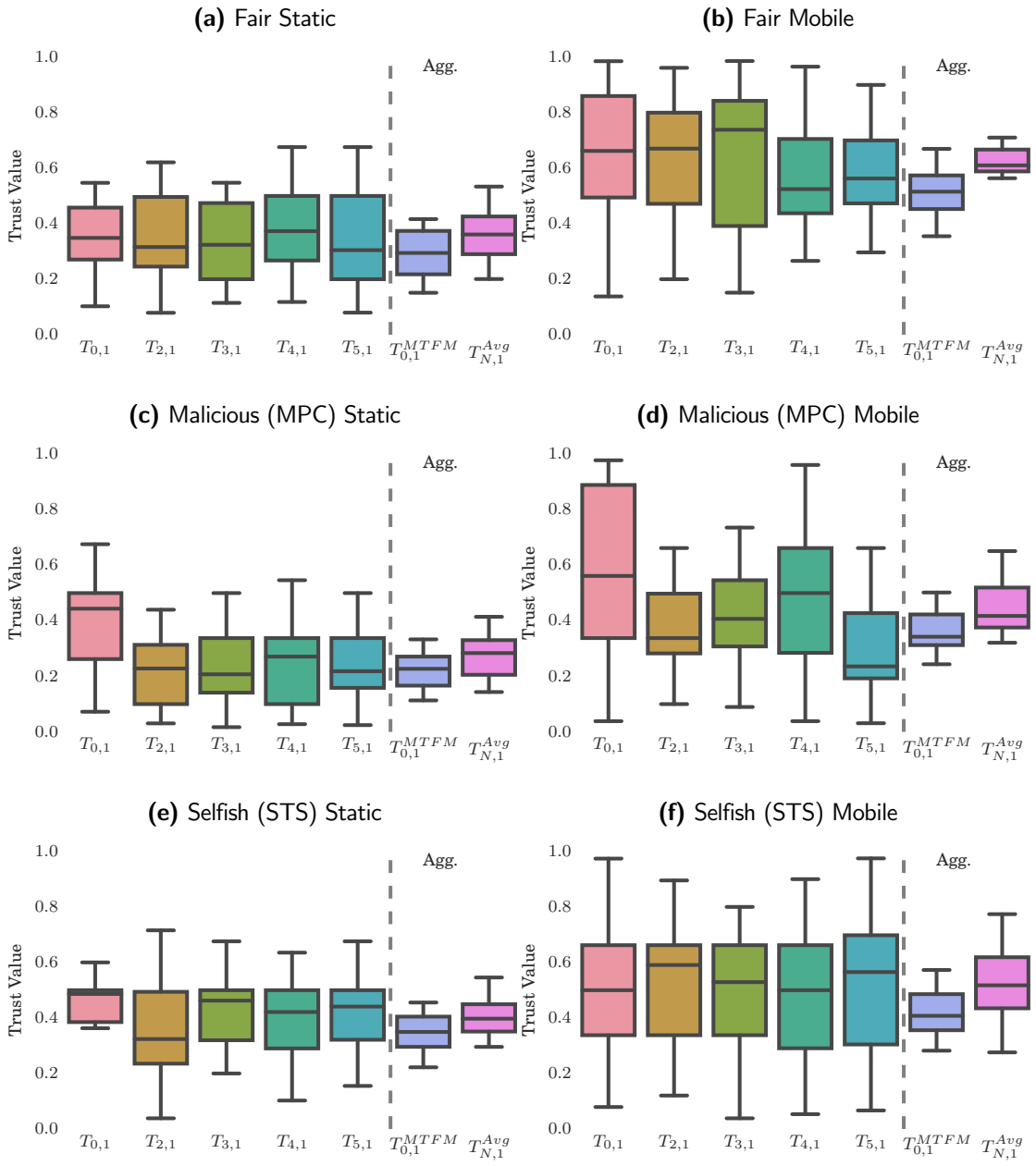
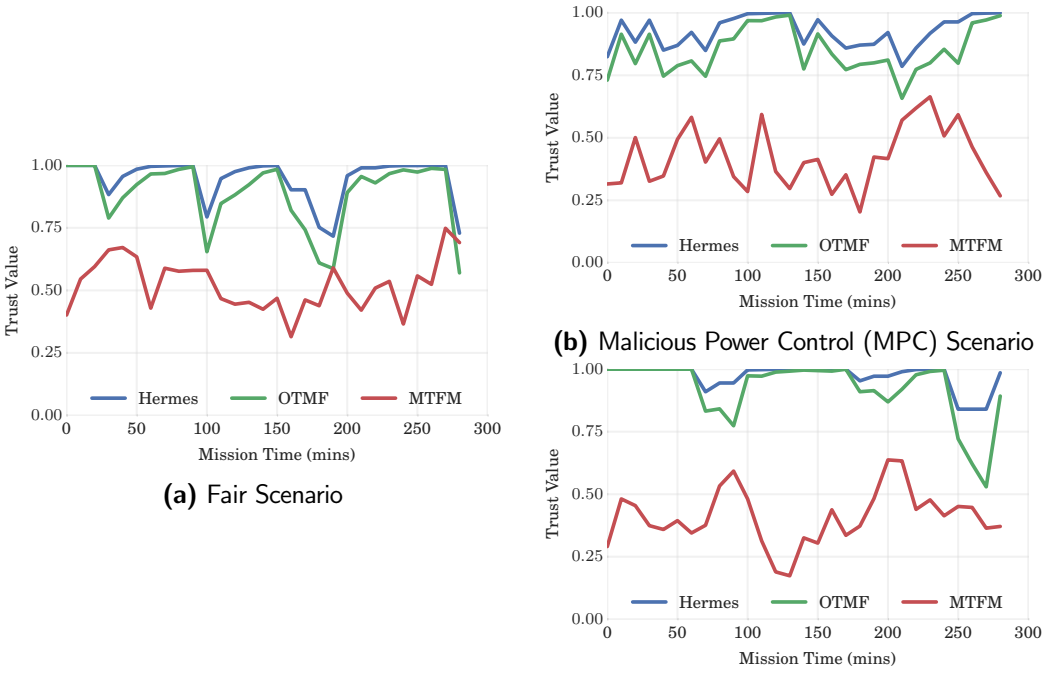


Figure 4.13: MTFM Trust assessments of n_1 ($T_{X,1}$), showing Direct, Recommender and Indirect relationships, and derived aggregates

only genuinely lost packets occur when a node moves completely out of range of any other node and time out occurs in route discovery rather than transmission (See **??**). As such, confirmed packet losses are relatively rare and in a delaying network like this, it is difficult to set a differentiating time out between packets that are in the network but queued, and packets that are actually "lost".

The single metric TMFs used in conventional MANETs require regular and constant input to shape and adjust their evaluations, which for a network with significant and irregular delays such as this, is not practical. This renders OTMF and Hermes assessment at best uninformative and at worst misleading; consistently providing nodes a high trust assessment as they have very little information to extract trust from.



(a) Fair Scenario

(b) Malicious Power Control (MPC) Scenario

(c) Selfish Target Selection (STS) Scenario

Figure 4.14: $T_{0,1}$ for Hermes, OTMF and MTFM assessment values for fair and malicious behaviours in the fully mobile scenario

?? shows a comparison between the unweighted response of MTFM compared to OTMF and Hermes assessment functions on the same data for the fair, malicious and selfish behaviours respectively. This time-series perspective demonstrates how noisy and variable the assessments from all TMFs in this environment. For clarity, ?? displays a compressed-time perspective, showing the overall trend and sensitivity of the different frameworks in the same scenarios as shown in ??. It is important to note a distinction between the expectations of MTFM compared to other TMFs; MTFM is primarily concerned with the identification of differences in the behaviours of nodes in a network, and is relative rather than absolute. That is to say that under MTFM, nodes are compared against the worst current performances across metrics of other observed nodes and graded against them, rather than the absolute (objective) approach taken by many TMFs.

In the case of the MPC In these cases, particularly since the methods of attack were not directly related to PLR, OTMF and Hermes have not registered significant activity in either misbehaviour when compared to the fair scenario. The difference between the MTFM trust assessments under "fair" and "malicious" behaviour is lowered by $\approx$10% in both cases, in terms of the mean values returned. At run time, similar results could be attained by an Exponentially Weighted Moving Average (EWMA).

On their own, neither OTMF, Hermes, or unbiased MTFM appear to be effective in detecting or identifying malicious behaviour in this environment, in fact OTMF and Hermes don't appear to differentiate between fair and selfish scenarios at all.

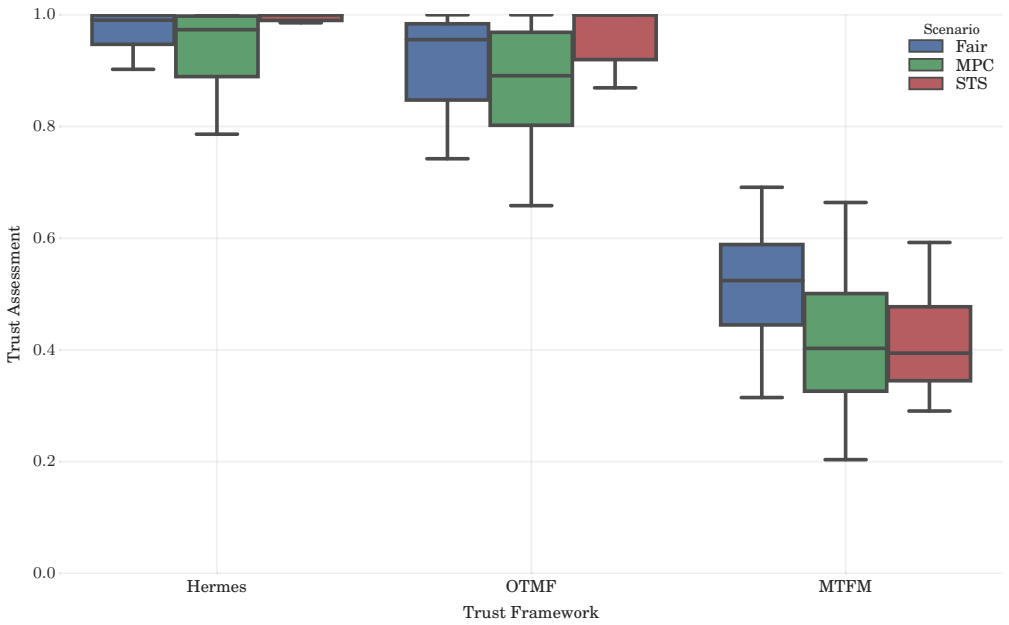


Figure 4.15: Alternative Visualisation of TMF performance comparison across Fair, MPC, and STS scenarios

4.4.2 Metric Vector Weighting

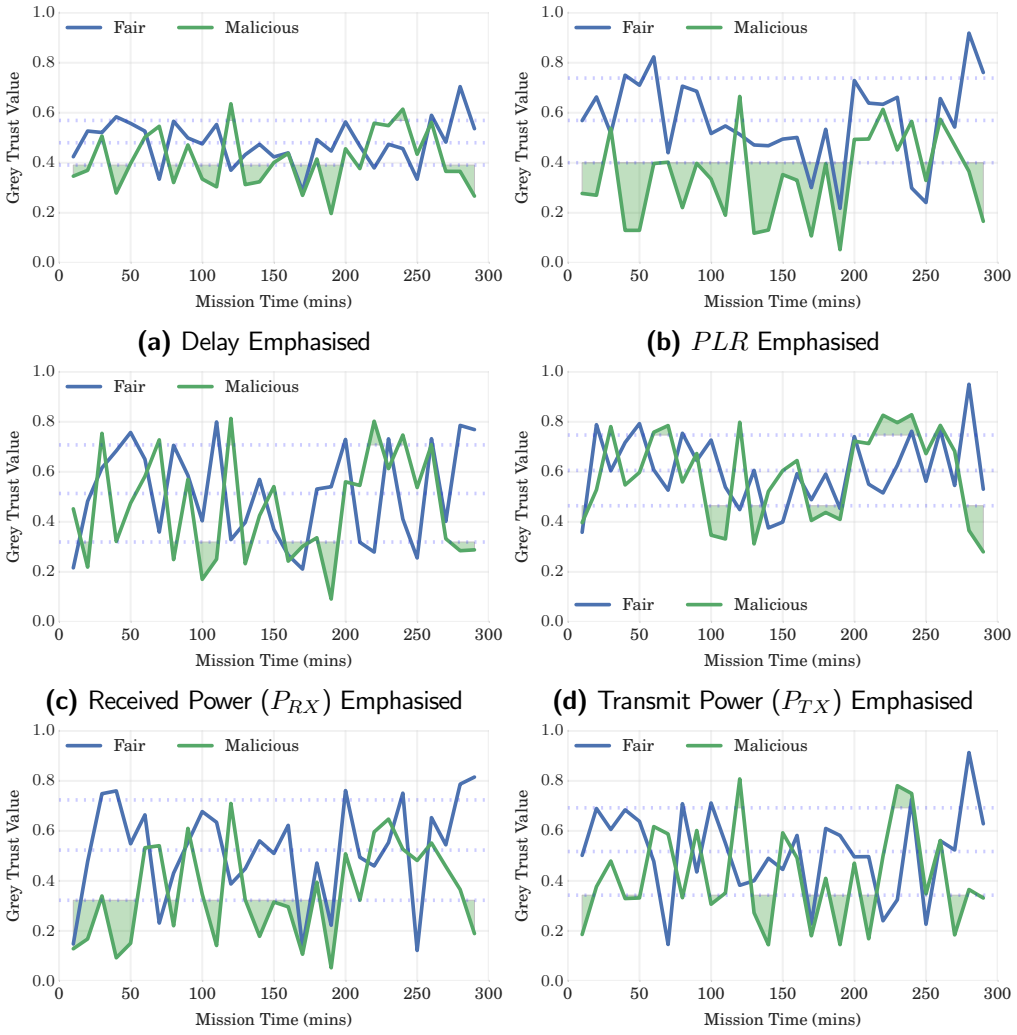
A sequence of vectors that preferentially weight each metric in (??) to each of the three simulation runs. For a metric weight vector H, where the metric m_j is emphasised as being twice as important as the other metrics, forming an initial weighting vector $H' = [h_i ... h_M]$ such that $h_i = 1 \forall i \neq j; h_j = 2$. That vector H' is normalised such that $\sum H = 1$ by $H = \frac{H'}{\sum H'}$. Using this process the primary aspects of an attack can be extracted and highlighted by comparing against the deviation from the "fair" result set.

Fig. ?? shows that the malicious node is consistently outside the $\pm\sigma$ (one standard deviation above and below the mean) envelope of the fair scenario it's being compared to. This is particularly true for PLR, with smaller impacts on delay, received power and offered load. This weighted delta in received throughput is minimal to insignificant compared to the width of the detection envelope, occasionally breaching the envelope for a short period.

In the selfish case (Fig. ??) a much lower weighted delta in PLR and delay is observed, with greatly increased impact on transmission power. In comparison to [?], these results are qualitatively similar, however here the differences between the fair case and the misbehaviours are less clear than in the comparable terrestrial space. ?] show similar types of behaviour but report a weighted delta from ≈ 0.4 to ≈ 0.9 across the simulation period, compared to our maximum delta in P_{TX} in selfish behaviour (Fig. ??) of ≈ 0.3 for an inconsistent interval.

4.4.3 Weight Significance Analysis for Behaviour Classification

For a more quantitative assessment of the viability of multi-metric trust assessment methods, taking the qualitative analysis above and apply a Random Forest regression [?] to assess the relative importance of the selected metrics on relative detectability of malicious behaviour.



(a) Delay Emphasised **(b)** PLR Emphasised

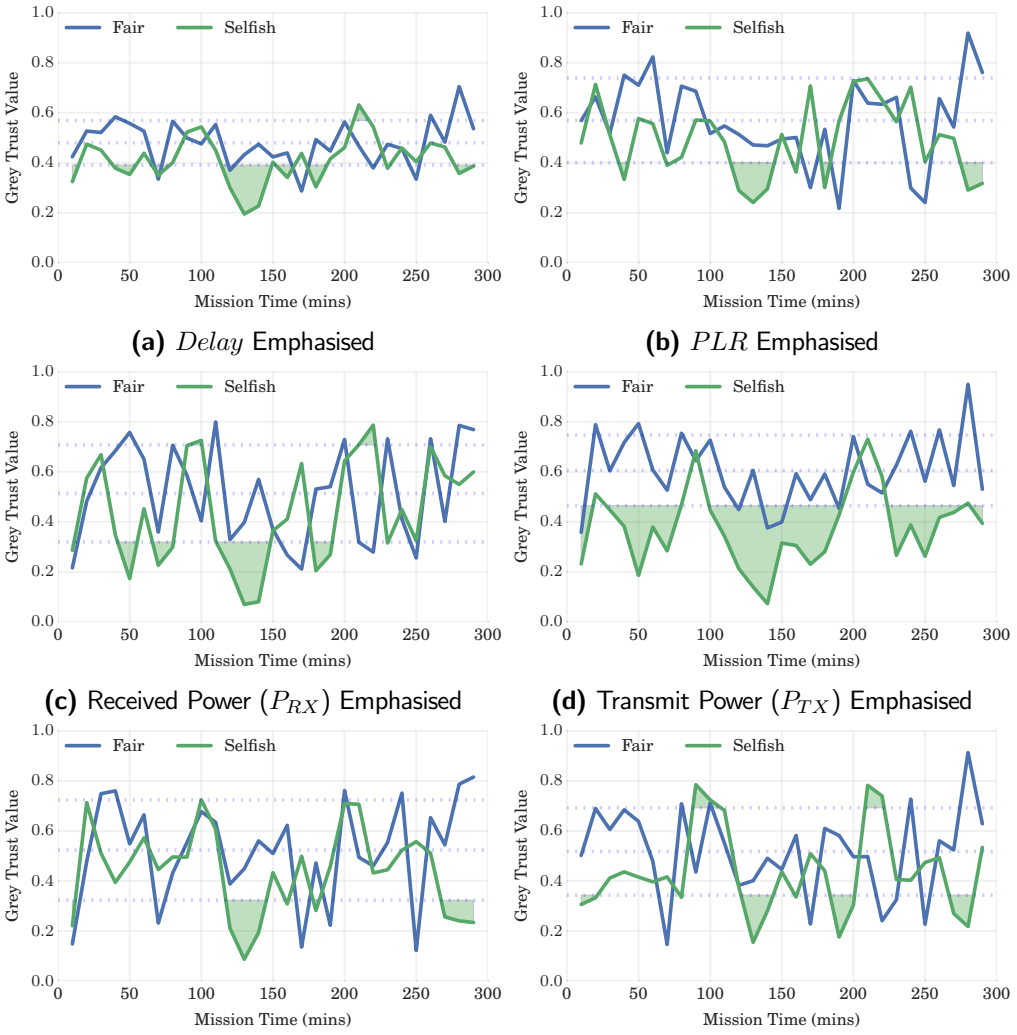
(c) Received Power (P_{RX}) Emphasised **(d)** Transmit Power (P_{TX}) Emphasised

(e) Throughput (S) Emphasised **(f)** Offered Load (G) Emphasised

Figure 4.16: $T_{1,0}^{\text{MTFM}}$ in the All Mobile case for the Malicious Power Control behaviour, including dashed $\pm\sigma$ envelope about the fair scenario

Random Forest accomplishes this by generating a large number of random regression trees and prunes these trees to fit incoming data. The target function for this regression was the area between the target behaviours weighted T_{MTFM} curve and the $\pm\sigma$ envelope of the base behaviour as shaded in Figs. **??** and **??**. From this training process, the relative importance of each input feature (metric) can be inferred in terms of how good it is to differentiate between the fair case and a given misbehaviour. Additionally a cross correlation analysis is performed to establish the correlations between given metric weighting emphasis and the output of the target function. Our intention is to establish the metrics that not only differentiate both misbehaviours from the fair case, but also what metrics differentiate the two misbehaviours from each other.

Applying this target regression to 729 different metric weight vector emphasis combinations reveals that each of the three combinations (i.e. comparing fair to misbehaviours, and comparing the misbehaviours) present distinct patterns of significance in three primary metrics; received throughput, transmitted power, and PLR, with delay, received power and transmitted throughput



(a) *Delay* Emphasised **(b)** PLR Emphasised

(c) Received Power (P_{RX}) Emphasised **(d)** Transmit Power (P_{TX}) Emphasised

(e) Throughput (S) Emphasised **(f)** Offered Load (G) Emphasised

Figure 4.17: $T_{1,0}^{\text{MTFM}}$ in the All Mobile case for the Selfish Target Selection behaviour, including dashed $\pm\sigma$ envelope about the fair scenario

playing a lesser role. Practically this means that in order to accurately distinguish between these scenarios, these primary metrics should be higher-weighted in the generation of $T_{1,MTFM}$ in (**??**).

It may initially appear odd that the relative significance of the received throughput is similar between all three scenario combinations, however a correlation analysis shows that in the MPC attack; the received throughput is positively correlated with successful classification against the fair case ($R = +0.71, p \approx 10^{-100}$), while the inverse is the case for the STS attack ($R = -0.70, p \approx 10^{-100}$). It is expected that Transmitted power should be the defining characteristic of STS ($R = +0.72, p < 10^{-100}$) as the node is acting fairly from a protocol perspective but is acting unfairly at a higher (incentive) level; it is performing fairly in terms of it's communications with other nodes, however it is preferring to communicate with nodes that it can expend less energy communicating with. A summary of these correlations is shown in Table. **??**.

Comparing Figs. **??**, **??**, and **??**, while it is possible that in a cleaner, less sparse, and less noisy environment, OTMF would be able to detect the MPC behaviour, Fig. **??** shows that PLR plays almost no part at all in detecting the STS behaviour, and so OTMF would not detect the attack.

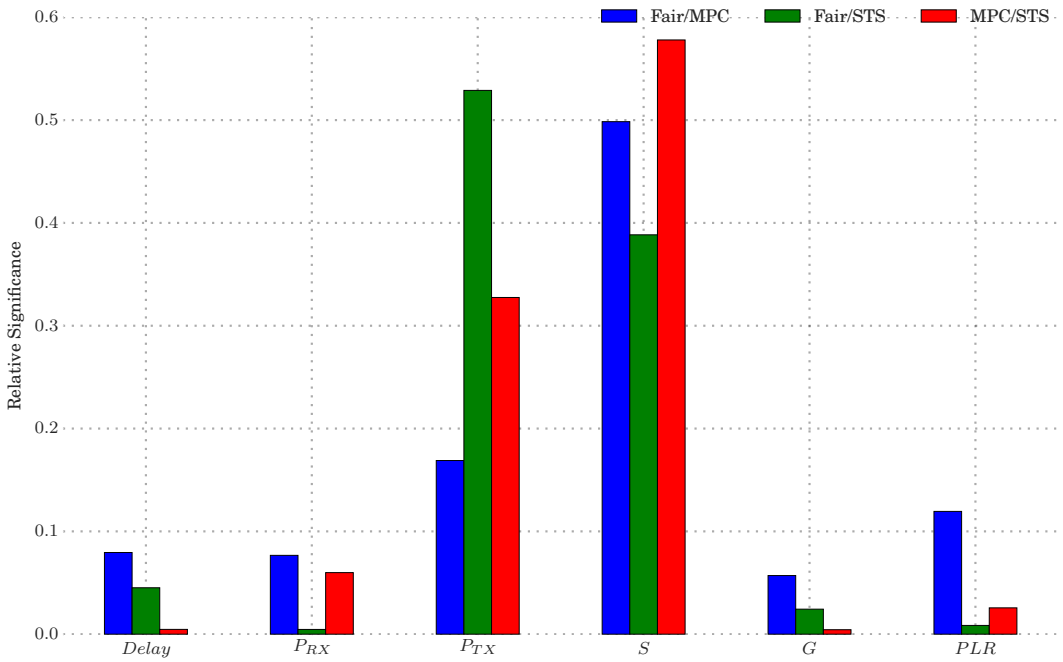


Figure 4.18: Random Forest Factor Analysis of Malicious (MPC, Selfish (STS) and Fair behaviours compared against each-other

Table 4.3: Correlation Coefficients between metric weights and behaviour detection targets

Correlation	Delay	P_{RX}	P_{TX}	G	S	PLR
Fair / MPC	0.199	0.159	-0.416	0.708	-0.238	-0.401
Fair / STS	0.179	-0.009	0.724	-0.697	-0.145	-0.052
MPC / STS	0.058	-0.134	0.146	-0.768	0.052	0.146

As such this presents the open opportunity to develop a heuristic weight search scheme to detect malicious behaviour without the comparison to the fair scenario. This would be accomplished by assessing the impact of differential metric weighting on the mean trust assessment rather than comparing co-weighted valuations across scenarios.

4.5 Conclusion

It has been demonstrated that existing MANET Trust Management Frameworks are not directly suitable to the sparse, noisy, and dynamic underwater medium. By comparing the operation and performance of trust establishment in MANETs in a simulated underwater environment has demonstrated that in order to have any reasonable expectation of performance, that throughput and delay responses must be characterised before implementing trust. It is shown that across the tested TMFs, Hermes and OTMF demonstrate a small discriminating factor between normal behaviour and one of the misbehaviours, this discrimination is not very clear and OTMF does

not detect the STS behaviour at all, while MTFM shows a significant ($\approx$10%) mean assessment drop the both misbehaving cases. In terms of behaviour discrimination, the MTFM value only displays a small immediate difference between the two misbehaviours, however it has been shown that by exploring the metric space by weight variation, the existence and nature of the malicious behaviour can be discovered. Another difference is that MTFM is significantly more computationally intensive than the relatively simple Hermes / OTMF algorithms. The repeated metric re-weighting required for real time behaviour detection is therefore an area that requires optimization. With significant delays (from seconds to many minutes), in a fading, refractive medium with varying propagation characteristics, the environment is not as predictable or performant as classical MANET TMF deployment environments.

It is shown that, without significant adaptation, single metric probabilistic estimation based TMFs are ineffective in such an environment. Additionally, it's clear that existing frameworks are overly optimistic about the nature and stability of the communications channel, and can overlook characteristics that are useful for assessing the behaviour of nodes in the network. This indicates that there is a good case, particularly within constrained MANETs as this, for multi-vector, and even multi-domain trust assessment, where metrics about the communications network and topology would be brought together with information about the physical behaviours and operations of nodes to assess trust.

A significant additional factor of trust assessment in such a constrained environment is that there may be long periods where two edge nodes (for instance, $n_0 \rightarrow n_5$) may not interact at all. This can be due to a range of factors beyond malicious behaviour, including simple random scheduling coincidence and intermediate or neighbouring nodes collectively causing long back-off or contention periods. This disconnection hinders trust assessment in two ways; assessing nodes that do not receive timely recommendations may make decisions based on very old data, and malicious nodes have a long dwelling time where they can operate under a reasonable certainty that the TMF will not detect it (especially if the node itself is behaving disruptively).

However the demonstrated noisiness and sparseness of communications metrics for trust assessment indicate that it may be more beneficial to look to other domains beyond communications to establish trust, such as the physical domain where we are concerned with the motion, placement and behaviour of the network nodes, which is investigated in the next chapter.

Chapter 5

Use of Physical Behaviours for Trust Assessment

5.1 Physical Behaviours for Trust

5.1.1 Physical Metrics

The aim of any TMF is to constrain the operation of a system such that any "trusted" behaviour is inherently "correct" behaviour; by monitoring PLR, delay, and throughput etc. In the communications domain, TMFs aim to optimise the efficiency of these aspects of the networks performance.

Looking at the physical domain, the question becomes "What characteristics of the nodes operations require constraint or optimisation, and what information is exposed to assess those?".

Fundamentally, the physical information available (or at least, is reasonable to assume) is simple positional and velocity information reported by the nodes itself. These assessments could be augmented with the use of sonar or visual tracking at short distances, or through time-of-flight positioning, but in the marine environment, both of these are difficult to accomplish and maintain consistently.

As for what characteristics of operations require optimisation, an assumption can be made that the primary threat is of a masquerading or damage of a node by a competent attacker, i.e. a physical TMF should identify nodes that are behaving oddly. Additionally, and in a less threatening manner, an additional aspect to optimise for is efficiency of mobility and communication, i.e. maintaining relative proximity to lower communications energy costs, and minimise expensive or "thrusty" course corrections.

Therefore, based on a fuzzy incomplete knowledge of the position and velocities of fleet/team members over time, designed around the REMUS 100 AUVs Kinematics, three primary initial metrics are arrived at:

1. *Inter-Node Distance Deviation (INDD)*, a second order measure of the variation in the average distances between nodes in a squad

2. *Inter-Node Heading Deviation (INHD)*, similar to Inter-Node Distance Deviation (INDD) but based on instantaneous unit velocity, i.e. the direction of travel

3. *Node Speed*, looking at the variability of through-the-water speed of each node in the squad with respect to the observable squads speeds.

Using these metrics, appropriate behaviour within a dynamic fleet can be assessed dynamically and in a decentralised fashion.

These physical metrics are used to encompass the relative distributions and activities of nodes within the network. As such, these metrics completely encapsulate and abstract the physical behaviour of any node, potentially performing any misbehaviour. Given that local nodes within the team are aware of the reported positions and velocities of their neighbours, it is believed that this is a reasonable set of metrics to establish the usefulness of physical metrics of trust assessment.

Additional metric constructions may be more suitable for certain contexts, platforms or operations, however these were selected in collaboration with UK DSTL and NATO CMRE as suitable, generic, assessments, viable on most current platforms in most current deployment schemes[1].

$$INDD_{i,j} = \frac{|P_j - \sum_x \frac{P_x}{N}|}{\frac{1}{N}\sum_x \sum_y |P_x - P_y| (\forall x \neq y)} \tag{5.1}$$

$$INHD_{i,j} = \hat{v}|v = V_j - \sum_x \frac{V_x}{N} \tag{5.2}$$

$$V_{i,j} = |V_j| \tag{5.3}$$

Where i and j are indices denoting the current observer node and the current observed node respectively; x is a summation index representing other nodes in the observers region of concern; P_j is the $[x,y,z]$ absolute position of the observed node (relative to some coordinated origin point agreed upon at launch) and V_j is the $[x,y,z]$ velocity of the observed node.

Thus, the metric vector used for the physical-trust assessment from one observer node to a given target node is;

$$X_{i,j} = \{\text{INDD}_{i,j}, \text{INHD}_{i,j},, V_{i,j}\} \tag{5.4}$$

At each time-step, each node will have a separate X assessment vector for each node it has observed in that time. Ergo the fleet or team as a whole will have $N \times N - 1$ assessment vectors at each timestep.

5.1.2 Physical Misbehaviours

Misbehaviours in the communications space is heavily investigated area in MANETs [**? ? ? ?**], but attacks and misbehaviours in the physical space are far less explored. Both in terrestrial and underwater contexts, as MANET applications expand and become increasingly *de rigueur*, the impacts of physical or operational misbehaviour become increasingly relevant. As in the communications

[1] An additionally prototyped metric was Reported Position Deviation which used a per-node Enhanced Kalman filter based "god view" estimator that constructed positional models based on highly accurate timing and time-of-flight modelling for non-linear channel paths to predict the movements of squad members and report discrepancies against their periodic positional updates (assumed to be part of a normal broadcast protocol), however this investigation is outside the scope of this current work

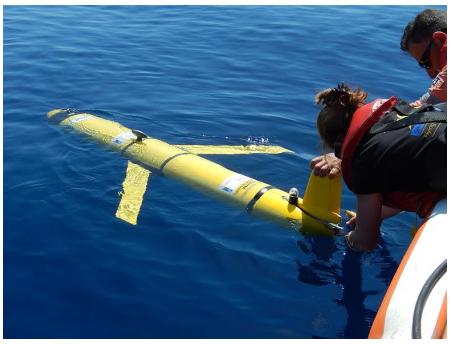


Figure 5.1: REMUS 100 Craft deployed at CMRE, La Spezia, Italy

space, the primary drivers of any "misbehaviour" come under two general categories; selfish operation or malicious subterfuge. Autonomous MANETs in general rely (or are at least, most effective) when all nodes operate fairly, be that in terms of their bandwidth sharing, energy usage, routing optimality or other factors. Physically, if a node is being "selfish", it may preferentially move to the edge of a network to minimise it's dynamic work allocation, or depending on it's intent, may insert itself into the centre of a network to maximise it's ability to capture, monitor, and manipulate traffic going across the network. In the context of a secure operation (or one that's assumed to be secure), there is also the opportunity for capturing a legitimate node and replacing it with a modified clone. Assuming a highly capable outside actor and a multi-channel communications opportunity, there is also the possibility of a node appearing to "play along" with the crowd that occasionally breaks rank to route internal transmissions to a outside agent. In the underwater context this may mean an AUV following the rest of a team along a survey path and occasionally "breaking surface" to communicate to a malicious controller using a secondary communications link such as WiFi or satcomms. Alternatively, if an inserted node is not totally aware of a given mission parameter, such as a particular survey or waypointing path, it may simply follow along, hoping not to be noticed.

In all these cases, such behaviour involves some element of behaving differently from the rest of the team, however, there are other cases where such individual "deviance" is observed; where a node is in some kind of mechanical "failure state". In the underwater context, this could be damage to the drive-train or navigation systems, causing it to lag behind or consistently drift off course. An ideal physical trust management system would be able to differentiate between both "malicious" behaviours and "failing" behaviours.

To investigate this hypothesis, we create two "bad" behaviours; one "malicious", where a cloned node is unaware of the missions' survey parameters and attempts to "hide" among the fleet, and a "failing" node, with an impaired drive train, increasing the drag force on the nodes movement. These two behaviours are designated *Shadow* and *SlowCoach* respectively.

5.2 Simulation and Validation

5.2.1 Simulation Background

Simulations were conducted using a Python based simulation framework, SimPy [**?**], with a network stack built upon AUVNetSim [**?**], with transmission parameters taken from and

Table 5.1: REMUS 100 Mobility Constraints as applied in simulation

Parameter	Unit	Value
Length	m	5.5
Diameter	m	0.5
Mass	kg	37
Max Speed	ms^{-1}	2.5
Cruising Speed	ms^{-1}	1.5
Max X-axis Turn	$^{\circ}s^{-1}$	4.5
Max Y-axis Turn	$^{\circ}s^{-1}$	4.5
Max Z-axis Turn	$^{\circ}s^{-1}$	4.5
Axial Drag Coefficient (c_d)	NA	3
Cross Section Area	m^2	0.13

validated against [**?**] and [**?**]. For the purposes of this chapter, this network is used for the dissemination of node location information, assuming suitable compression of internally assumed location data compressed into one 4096 bit acoustic data frame, with the network overall emitting approximately 10 frames a minute. Node kinematics are modelled on REMUS 100 AUVs (**??**), based on limits and core characteristics given in [**? ? ?**].[2] These limits are given in Table **??**.

5.2.2 Node Control Modelling

In our investigation, we use the example of a Port Protection scenario, where a team of six AUVs are tasked with surveying a simplified harbour; in this case a 1kmx1kmx100m cuboid volume. This is accomplished through a distributed way point system where by the team overall must "check" several points around the exterior and interior of this volume in reasonable time. In addition to this, there is a reasoned requirement for both collision avoidance and a pressure for the fleet to maintain communications distance.

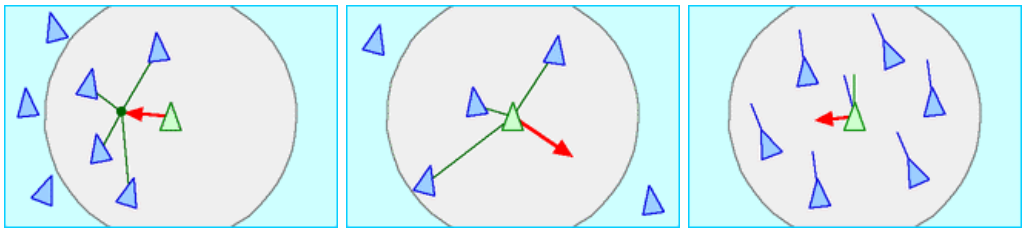
These are encapsulated as three heuristic rules; Cohesion, Repulsion and Alignment.

$$F_{j,C} = F_{+}\left(p_j, \frac{1}{N}\sum_{\forall i \neq j}^{N} p_i, d_{max}\right) \tag{5.5}$$

$$F_{j,R} = \sum_{\forall i \neq j}^{N} F_{-}\left(p_j, p_i, d_{max}\right) \Big| d_{max} > \|p_i - p_j\|\Big) \tag{5.6}$$

$$F_{j,A} = \frac{1}{N} \cdot \left(\sum_{\forall i \neq j}^{N} \hat{v_i}\right) \tag{5.7}$$

[2] While the hydrodynamics of the control surfaces of the AUVs are not modelled in this case, axial drag is modelled as a resistive inertial force on the craft.



(a) Cohesion **(b)** Repulsion **(c)** Alignment

Figure 5.2: Visual representation of the basic Boidean collision avoidance rules used

Where F's are force-vectors applied to the internal guidance of the AUV, $F_{j,C}$ representing Cohesion, $F_{j,R}$ representing Repulsion, and $F_{j,A}$ as Alignment: F_+ is a scaled vector attraction function, and F_- is an equivalent repulsion function

$$F_+(p^a,p^i) = (\widehat{p^a - p^i}) \times \frac{|p^a - p^i|}{d} \tag{5.8}$$

$$F_-(p^r,p^i) = (\widehat{p^i - p^r}) \times \frac{|p^r - p^i|}{d} \tag{5.9}$$

In essence, the fleet is simultaneously attracted to its current target waypoint as well as a lesser attraction to the centre of the fleet to retain communications.

5.2.3 Standards of Accuracy

The key question of this chapter is to assess the advantages and disadvantages of utilising trust from the physical domain.

It is important to clarify what is meant by "effective" in this case; the "effectiveness" of any trust assessment framework is taken as consisting of several parts, the *accuracy* of detection and identification of a particular misbehaviour, the *complexity* of such analysis, including any specific training required, and the *differentiability* of behaviours using given metrics.

In this case we are particularly interested in the accuracy of detection and identification of malicious / failing behaviours, and as such are looking at three key characteristics of accuracy; true detection accuracy (what percentage of "bad" behaviours are detected at all); false positive rates (what percentage of "control" behaviours are detected as being "bad"); and misidentification rates (how many instances of one bad behaviour are mischaracterised as the other and vice versa.

As such we have three primary questions to answer to establish if these metrics are useful: How accurate are these metrics in being able to easily differentiate between Normal and Abnormal behaviours in terms of True-Positive and False-Positive rates? What differentiation of response, if any, is there between the stated abnormal behaviours? Can a simple classification be built to characterise these differentiations of response, and what is it's True-Positive/False-Positive accuracy?

5.2.4 Analysis

Having established the metrics under investigation, 64 simulation runs are executed for each scenario (i.e. one node "Maliciously" following the fleet with no mission information (Shadow),

one "Failing" node with simulated drive train issues (Shadow), and one baseline control scenario where all nodes are behaving appropriately (Control). Each of these simulated missions last for an hour, matching realistic deployment times based on current MOD/NATO operations[**?**].

Metric Cleaning

In order to assess the viability of using the previously discussed metrics, the raw motion paths recorded by the simulation are fed into an analysis pipeline aimed at abstracting the instantaneous observed values into derived deviations from "normal" behaviour in the team.

$$d_{i,j}^{m,t} = x_{i,j}^{m,t} - \frac{\sum_k x_{i,k}^{m,t}}{|M|} \tag{5.10}$$

$$\alpha_{i,j}^{m,t} = |\frac{d_{i,j}^{m,t}}{\sigma d_{i,j}^{m,t}}| \tag{5.11}$$

Where i and j are indices denoting the current observer node and the current observed node respectively; x is a summation index representing other nodes in the observers region of concern; X is the vector of metrics from **??**; d is an intermediate value of the distance of a given observation from the mean, and α is a resulting normalised response value in terms of it's deviation from the mean.

Behaviour Detection and Classification

A simple misbehaviour detection is to apply Dixon's Q-test [**?**] to the resultant $\sum\alpha$ values for each node for each metric for each run establishing if a "misbehaving node" exists in a given run, and if so, attempt to identify that misbehaving node. For our initial investigation we will use a Confidence Interval of 95%.

Our initial hypothesis is that by using observations of the previously stated physical metrics, that we will be able to detect and identify misbehaviours. Within that context, this Confidence Interval indicates that we would expect only a 5% chance that any run or node identified using the Q-test to *not* be a misbehaving run/node. Further, due to the range of metrics available, by applying the Q-test on a per-metric basis, we can use the "votes" of each metric as a simplified consensus classifier. This classifier may allow us to characterise some aspect of a given misbehaviour in terms of metrics it heavily impacts, and those that are less affected, finding some differentiating-limit between certain behaviours using certain metrics.

Operational Performance Metrics

While not the focus of this chapter, we are also concerned with the impact of these misbehaviours on the mission efficiency of the team overall. We monitor this in three main measurements; the "speed" of the fleet in terms of how many of it's port-protection way points it successfully approaches and passes, the total energy used for communications, and the average end-to-end delay in the acoustic network. We would expect that any misbehaviour in positioning will incur

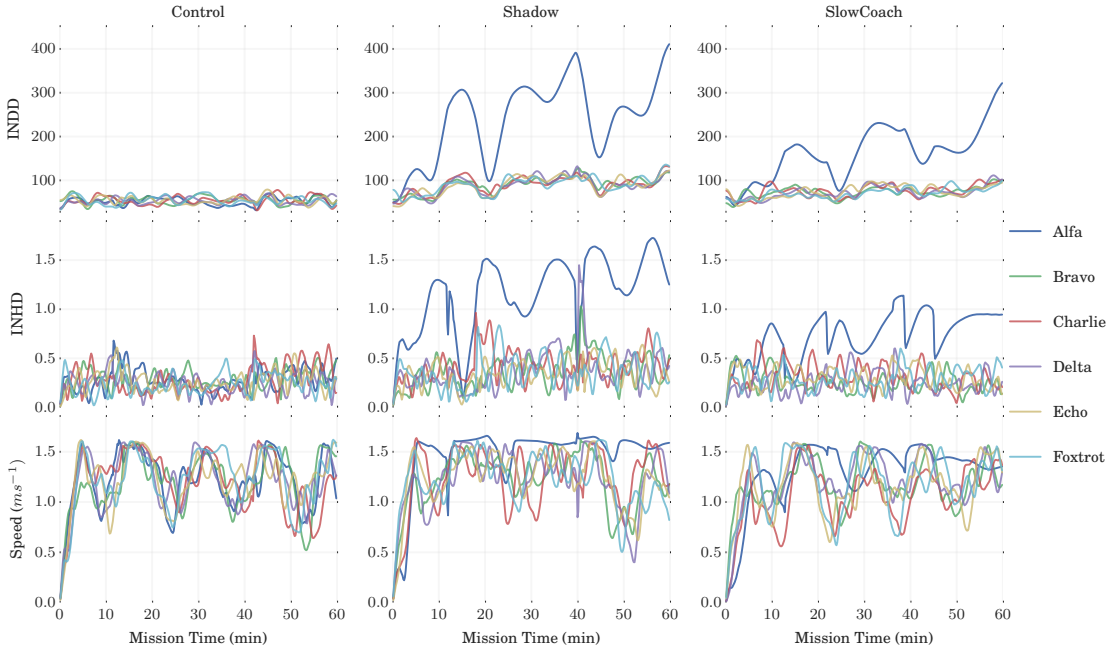


Figure 5.3: Observed Metric Values for one simulation of each behaviour ($x_{i,j}^{m,t}$ from (??))

some loss of efficiency, whether it is the fleet being slowed down by a straggler attempting to catch up or of a node moving in an unexpected fashion dragging the team temporarily off course. Given that in acoustic communications, transmission is energetically expensive while reception is not, and while physical misbehaviours will not impact the amount of offered load on the network, collisions induced by uneven distribution of nodes should have a small but measurable effect on energy used for packet reception.

5.3 Results and Discussion

Fig. ?? shows the raw metric values (vertically) from one run of each behaviour (horizontally), starting with the Control case, where all node are behaving properly with Alfa as the misbehaving node in the remaining cases. It clear that using the (unitless) INDD and Inter-Node Heading Deviation (INHD) metrics, Alfa is the outlier and other, fairly behaving, nodes are all consistent in their metric values. This outlier-response is not nearly as clear in the Speed metric case (bottom row of Fig. ??). This would be expected considering the cumulative factor of increasing distance between nodes if a given node is "lagging" behind.

From a behaviour-perspective, it appears that the Shadow behaviour is creating the largest, most obvious deviations.

In Fig. ?? the metric values are normalised as per (??). This has highlighted the outlying-characteristic of INDD and INHD; largely eliminating the other nodes-responses. In the Speed response of Fig. ??, the Speed metric is not obviously highlighting any significant misbehaviours in that metric.

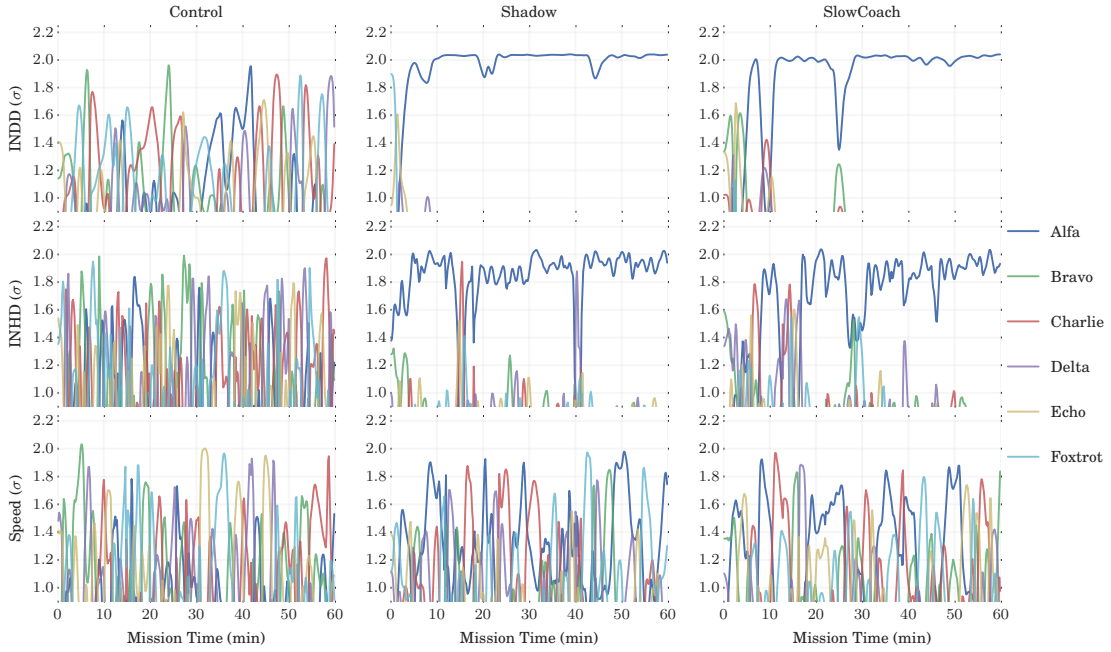


Figure 5.4: Normalised Deviance values from one simulation of each behaviour ($\alpha_{i,j}^{m,t}$ from (??))

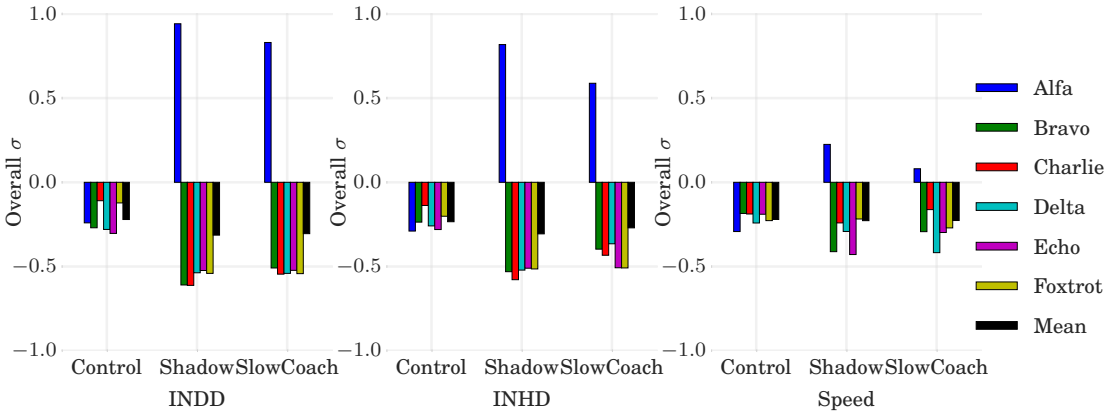


Figure 5.5: Per-Node-Per-Run deviance for each metric, normalised in time ($\sum\alpha/T$)

From Fig. **??**, it appears that Speed is being significantly affected by the differing behaviours, but much less so than INDD/INHD.

5.3.1 Detection of Misbehaviours

It has been demonstrated by graphical result that from the initial metrics set, INDD and INHD do appear to accurately and obviously identify the malicious node in the case that there is one. Using the deviance normalisation presented in (**??**), clear, almost contiguous areas under the Alfa-values are observed in Fig. **??** in the Shadow and SlowCoach misbehaviours. Further, from Fig. **??**, it is shown that while it is nowhere near as "clear" as the deviance in INDD and INHD, that the Speed metric is still registering a statistically significant deviation in both

Table 5.2: Overall Q-Test Outlier Correct Detection Accuracy

Behaviour	Mean	Std.
Control	0.927	0.261
Shadow	0.979	0.144
SlowCoach	0.792	0.408

Table 5.3: Per-Metric Q-Test Outlier Detection Accuracy

	Behaviour	INDD	INHD	Speed
Mean	Control	0.875	0.938	0.969
	Shadow	1.000	1.000	0.938
	SlowCoach	1.000	1.000	0.375
Std	Control	0.336	0.246	0.177
	Shadow	0.000	0.000	0.246
	SlowCoach	0.000	0.000	0.492

misbehaviours, and that the difference between the deviances in Speed may indicate a way to analytically differentiate between the two misbehaviours.

To investigate how this would relate to the ability to blindly detect misbehaviours, the Q-test is applied to $\Sigma\alpha$ results as used in Fig. **??**, to attempt to correctly establish:

1. if a node is misbehaving and
2. which node is misbehaving

As such, the "correctness" rule for assessing this strategy is that, in misbehaving cases, the Q-tests should return Alfa (otherwise a "Fail" is recorded), and in the Control case, the Q-test should assert that there are no obvious outliers, (otherwise a "Fail" is recorded again). In Table **??**, the Null case (Control behaviour) is correctly identified 92% of the time. The "malicious", Shadow misbehaviour is detected and identified 98% of the time, and the "failing", SlowCoach misbehaviour is identified just 79% of the time. These values match our intuition from Figs. **??** and **??**.

We can investigate this further by looking at the "correctness" of the assessments of each metric individually (Table **??**). In both misbehaviours, INDD and INHD correctly identify Alfa as the misbehaver 100% of the time. However, they misidentify a potential misbehaviour in the Control case 13% and 7% of the time respectively. Meanwhile, Speed correctly identified the Control case 97% of the time, and the Shadow case 94% of the time, but missed the SlowCoach behaviour 63% of the time. This result is surprising on the face of it, as SlowCoach is a misbehaviour that is exclusively about individual node speed and conceptually should have had a much larger impact on the simple Speed metric. However, the collaborative nature of the collision avoidance system, and the existing limits on node kinematics from Table **??** are masking this impact.

5.3.2 Identification of Misbehaviours

Having established the ability of INDD, INHD and Speed to all detect physical misbehaviour to a statistically significant level, and having shown that there is a demonstrable difference in response to different misbehaviours, we return to the last question from Sec. **??**; can a simple classifier based on a subset of our results be constructed, and can it be blindly applied to a new set of results successfully?

From (**??**), the per-metric-per-behaviour "Confidence" in the relationship between a given metric deviance and each behaviour is established. It is hypothesised that this confidence can be used as a signature for that metric.

$$C_i^m = \Sigma_t \sigma_i^m * \frac{N-1}{\sum_{x \neq i} \Sigma_t \sigma_x^m} \tag{5.12}$$

Table 5.4: Metric Confidence Responses for known behaviours (**??**)

	Behaviour	INDD	INHD	Speed
Mean	Control	1.064	0.966	1.010
	Shadow	4.059	3.374	2.098
	SlowCoach	4.246	3.352	1.491
Std	Control	0.262	0.113	0.132
	Shadow	0.398	0.436	0.206
	SlowCoach	0.198	0.288	0.180

Having demonstrated that the Null case (All nodes behaving fairly) can be identified to a strong degree of accuracy, our classifier will continue to use the Q-test across all metrics for that case and concentrate of differentiating the Shadow and SlowCoach behaviours where they exist.

From Table **??** it is clear that INDD and INHD have similar responses to both misbehaviours, with significant standard deviations, but the response of the Speed metric is much more stable and discernible; across the range of training simulation runs. In the SlowCoach behaviour, this Speed response centres around 1.5, while the Shadow behaviour centres around 2.0, with these centres being at least one standard deviation away from each other respectively.

Our generated classifier is formalised in (**??**).

$$C \rightarrow \begin{cases} Q^{95}(X) = \emptyset, & \text{Control} \\ Q^{95}(X) \neq \emptyset \wedge \text{Speed}^X \leq 1.75, & \text{Shadow} \\ Q^{95}(X) \neq \emptyset \wedge \text{Speed}^X > 1.75, & \text{SlowCoach} \end{cases} \tag{5.13}$$

Applying this simplified classifier to a blind test set of simulations (of the same scale) gives surprisingly positive results as shown in Table **??**, with greater than 90% identification rates for both misbehaviours. However, in the Null (Control) case we experience a false-positive rate of nearly 30%, that is to say that in the case where there is no misbehaviour, 30% of the time a node will be misidentified as misbehaving when it is not.

Table 5.5: Successful Identification rates on untrained results using (??)

True Behaviour	Probability of Correct Blind Identification
Control	0.719
Shadow	0.906
SlowCoach	0.938

If the rules of this classifier are loosened to require two deviating $Q^{95}(X)$ observations for detection of non-Control behaviours (i.e. multi-metric deviations), this false positive rate is eliminated while maintaining true-positive detection characteristics for misbehaviours.

$$C \rightarrow \begin{cases} |Q^{95}(X)| \leq 1, & \text{Control} \\ |Q^{95}(X)| > 1 \wedge \text{Speed}^X \leq 1.75, & \text{Shadow} \\ |Q^{95}(X)| > 1 \wedge \text{Speed}^X > 1.75, & \text{SlowCoach} \end{cases} \tag{5.14}$$

Table 5.6: Successful Identification rates on untrained results using (??), with outlier consensus checks

True Behaviour	Probability of Correct Blind Identification
Control	1.000
Shadow	0.906
SlowCoach	0.938

Given the simplicity of the applied classifiers, these are strongly positive results for the use of physical metrics for behaviour discrimination; with INDD and INHD proving as strong and obvious "canaries" of misbehaviour, and Speed in this case proving a capable differentiator between conceptually close misbehaviours.

5.3.3 Impacts of Misbehaviour on operational performance

The anticipated "small but measurable" effects to communications performance and energy usage are indeed extremely small and within the bounds of statistical uncertainty. One observation of merit was an observed 10% increase in end-to-end delay in the case of the Shadow behaviour; this is due to the misbehaving node "overshooting" the mission way points and thus temporarily looking local connection to nodes on the opposite side of the fleet from it, causing retransmissions thus, delays. As for physical efficiency, achievement rates were identical to within 2% error on each run across all behaviours, and fleet distance varied by a similar margin. It's possible that our selected behaviours were too unambitious in our impacts, and future work will have to investigate the impact of "heavy-handed" or destructive behaviours on the operational efficiency of autonomous networks.

5.4 Conclusion

In this chapter we have demonstrated that with current and on-the-horizon underwater localisation techniques, that in certain mobility models, that a set of relatively simple geometric abstractions (INDD, INHD, and Speed), between nodes as part of an Underwater MANET can be used as a Trust Assessment and Establishment metric.

These metrics are application-agnostic and could potentially be applied in other areas of mobile autonomy such as AUV operations and Autonomous Vehicular Networks.

We show, using a Port-Protection way point scenario built upon a Boidian collision prevention behaviour that in a simulated underwater environment, the outputs of these metrics can be used to detect and differentiate between exemplar malicious behaviour and potential failure states.

This verification further supports the assertions the authors have made previously that it is practical to extend Trust protocols such as MTFM [**?**] to include metrics and observations from the physical domain as well as those from the communication domain[**?**]. In the next chapter, this combination of physical and "logical" information is explored to establish if it further supports the decentralised and distributed establishment of observation based Trust.

Chapter 6

Multi-Domain Trust Assessment in Collaborative Marine MANETs

6.1 Initial Optimisation of Multi-Domain Trust with Predefined Domains

A key question in this chapter is to assess the advantages and disadvantages of utilising trust from across domains for MTFM. This includes a secondary question as to how trust assessments from these domains are most effectively combined or synthesised.

It is important to clarify what is meant by "effective" in this case; the "effectiveness" of any trust assessment framework is taken as consisting of several parts.

1. the *accuracy* of detection and identification of a particular misbehaviour
2. the *timeliness* of such detections
3. the *complexity* of such analysis, including any specific training required
4. the *commonality* of the results of any detections between perspectives (also termed "isomorphism" of results)

6.1.1 Communications Trust Metrics

The metric vector is constructed using those trust metrics that are applicable to the marine environment from [**?**], as the simulated marine acoustic modem stack does not operate on the same tiered data-rate approach as used in the 802.11 stack, the data rate metric was not included. Remaining metrics are; Delay, Received and Transmitted power, Throughput (S), Offered Load (G) and PLR.

Thus, the metric vector used for communications-trust assessment is;

$$X_{comms} = \{D, P_{RX}, P_{TX}, S, G, PLR\} \tag{6.1}$$

6.1.2 Physical Trust Metrics

From **??**; Three physical metrics are selected to encompass the relative distributions and activities of nodes within the network; INDD, INHD, and Node Speed. These metrics encapsulate the relative distributions of position and velocity within the fleet, optimising for the detection of outlying or deviant behaviour within the fleet.

Conceptually, INDD is a measure of the average spacing of an observed node with respect to its neighbours. INHD is a similar approach with respect to node orientation.

$$INDD_{i,j} = \frac{|P_j - \sum_x \frac{P_x}{N}|}{\frac{1}{N}\sum_x \sum_y |P_x - P_y| (\forall x \neq y)} \tag{6.2}$$

$$INHD_{i,j} = \hat{v}|v = V_j - \sum_x \frac{V_x}{N} \tag{6.3}$$

$$V_{i,j} = |V_j| \tag{6.4}$$

Thus, the metric vector used for physical-trust assessment is;

$$X_{phy} = \{\text{INDD}, \text{INHD}, V\} \tag{6.5}$$

6.1.3 Cross Domain Trust Metrics

This simplest possible combination is a vector concatenation across domain metric vectors; in this case;

$$X_{merge} = (X_{comms}|X_{phy}) = \{D, P_{RX}, P_{TX}, S, G, PLR, \text{INDD}, \text{INHD}, V\} \tag{6.6}$$

6.1.4 Metric Weight Analysis Scheme

From (**??**), the final trust values arrived at using MTFM are dependent on metric values, the weights assigned to each metric, and the structure of the g, b comparison vectors.

This permits the assessment of the significance of different metrics in the detection and identification of different behaviours. The primary aspects of a (mis)behaviour can be detected and assessed by comparing a weighted trust assessment against the deviation from a "fair" result set using the same weight, i.e. we are interested in the weight schemes that create the largest difference between fair and misbehaving cases.

For a metric weight vector H, where the metric m_j is emphasised as being twice as important as the other metrics, an initial weighting vector $H' = [h_i \cdots h_M]$ is formed such that $h_i = 1 \forall i \neq j; h_j = 2$. That vector H' is then scaled such that $\sum H = 1$ by $H = \frac{H'}{\sum H'}$.

The construction of the g and b vectors from (**??**) depends on the particular metric, e.g. Throughput (S) on a link is assumed to be positively correlated to trustworthiness and so follows the default construction ($g(S) \mapsto \max, b(S) \mapsto \min$), whereas in the case of a metric

such as delay, this relationship is inverted, i.e. longer delays indicate less trustworthy activity ($g(D) \mapsto \min, b(D) \mapsto \max$). This inversion relationship (i.e. those with the construction $g(x) \mapsto \min, b(x) \mapsto \max$) is signified by a negative weight.

In complex environments, the relationship between metrics trustworthiness correlations is not always as obvious as the throughput / delay examples. This phenomenon was mentioned by ?], but was manually configured for each metric for each behaviour and no analytical method for quantitatively establishing such relationships has been presented since.

With the nine selected metrics from across communications and physical behaviours, we can explore this metric space by varying the weights associated with each metric, and choose to emphasise across three levels; i.e. metrics can be ignored or over-emphasised. Naively this results in $3^9 = 19683$ combinations, however as these weights are being normalised, redundant duplicates can be eliminated, e.g. $[0,0,0,0,1,0,0,0,0] \equiv [0,0,0,0,2,0,0,0,0]$ leaving 18661 unique weights for analysis.

To assess the performance of a given weight combination (i.e. an optimisation factor), we are initially interested in the metric weight vector that consistently provides the largest deviation in the final trust value T across the cohort, i.e. producing the most clear detection of a node misbehaving in that particular fashion. This is approached as an inverse outlier filtering problem, and the range outside a $\pm\sigma$ envelope compared to the equivalent weighting in a known "fair" behaviour is selected to assess detection (or comparing to other misbehaviours to assess discrimination). See **??**. Note that at this point we establish "signatures" of different behaviours rather than optimal detection weights.

We apply a Random Forest regression [?] to assess the relative importance of the selected metrics on relative detectability of malicious behaviour. Random Forest accomplishes this by generating a large number of random regression trees and prune these trees based on how accurate they are in correctly matching the input data. In this case that data is the deviation in trust observed (ΔT) between a two behaviours, i.e. maximising the ability to tell the difference between two given behaviours (i.e. "Fair" and "Malicious"). A major advantage of Random Forest in this case is that by walking the most successful regression trees, we can acquire an already normalised maximal activation weight for the particular behaviour comparison being tested.

After establishing the importance of weights in particular behaviours, a final weight is arrived at by algorithmically those few metrics that are important, rather than having to further explore the computationally expensive weight-space.

Using this approach, the results of these simulations can be explored, condensing the multi-dimensional problem (target / observer / behaviour / metric / time) down to a more manageable level for analysis.

6.1.5 Significance Analysis

First the results of the Random Forest regression assessment are discussed; Figs **??** and **??**, show the resultant feature significances for Comms-only and Physical-only metric selections respectively, and in **??**, these metric spaces are brought together and reassessed.

In both single-domain cases, there are clear "signatures" in misbehaviours that don't directly target that domain (P_{RX} in the Physical Shadow and Slowcoach behaviours in Fig **??** and

Table 6.1: Multi Domain Metric Feature Correlation (X_{merge})

Misbehaviour	*Delay*	P_{RX}	P_{TX}	*S*	*PLR*	*G*	*INDD*	*INHD*	*Speed*
MPC	-0.187	0.129	0.579	0.006	0.069	-0.146	0.040	-0.190	-0.297
STS	-0.195	-0.035	0.019	-0.100	0.019	0.381	-0.209	0.057	0.062
Shadow	0.004	-0.654	0.030	-0.016	0.030	0.063	0.120	0.158	0.266
SlowCoach	-0.157	-0.533	0.013	-0.132	0.013	-0.028	0.159	0.206	0.460

INDD in the Selfish Target Selection behaviour in Fig **??**). This inter-domain activity is to be expected in MANETs in general, where the physical reality of the network (i.e. distance between nodes) directly impacts the behaviour of the logical communications network (i.e. delay between nodes), and is a useful characteristic coupling for differentiating potential misbehaviours.

?? attempts to lay out the results of Feature Extraction across a spectrum of Communications / Physical domain assessment, matching the assumed domains of the stated misbehaviours based on the resultant significance in relevant domain misbehaviours. From this, two alternative "domains" can be constructed; "Comms. Alt." and "Phys. Alt.", as artificial constructions of the relevant domains attempting to encapsulate the most responsive features for each misbehaviour-domain. The results from this domain grouping are shown along side the Full, Comms, and Phys domains.

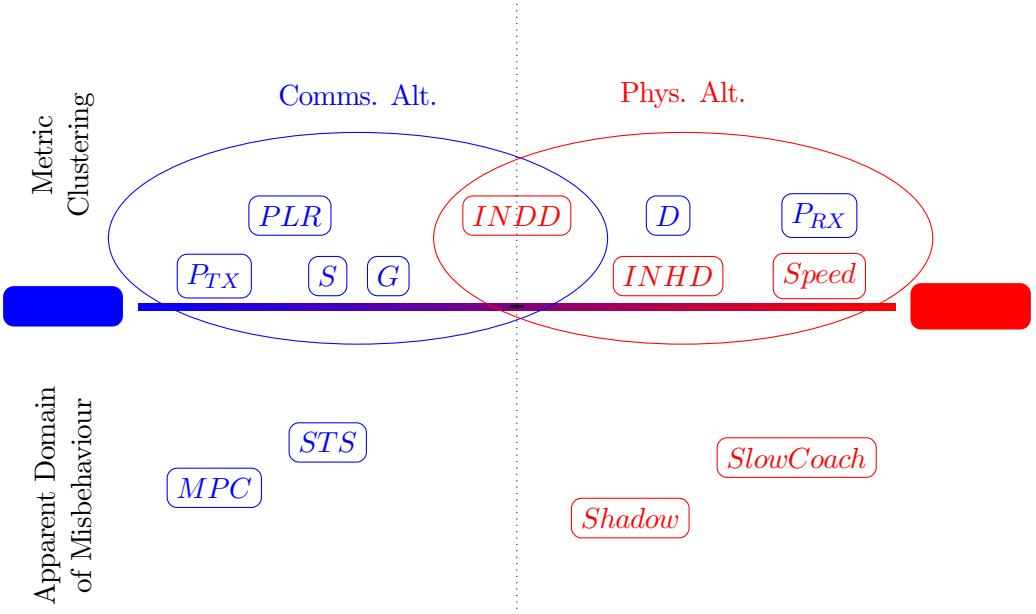


Figure 6.1: Assumptions made about the relevant domains of impact / detectability of misbehaviours, and domain relevance of metrics, may not be optimal

6.1.6 Weight Assessment

From this significance information, a "estimated" signature for each behaviour can be inferred, which can then be fed back into MTFM. The aim of this iteration is to minimise the number of

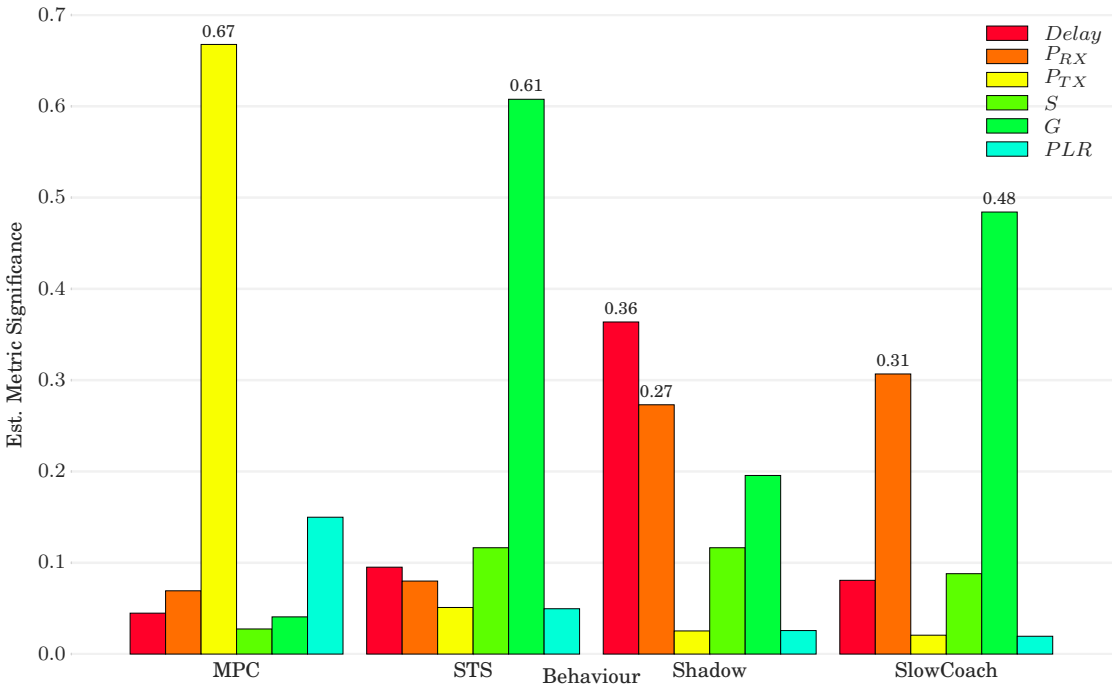


Figure 6.2: Communications Metric Features (X_{comms})

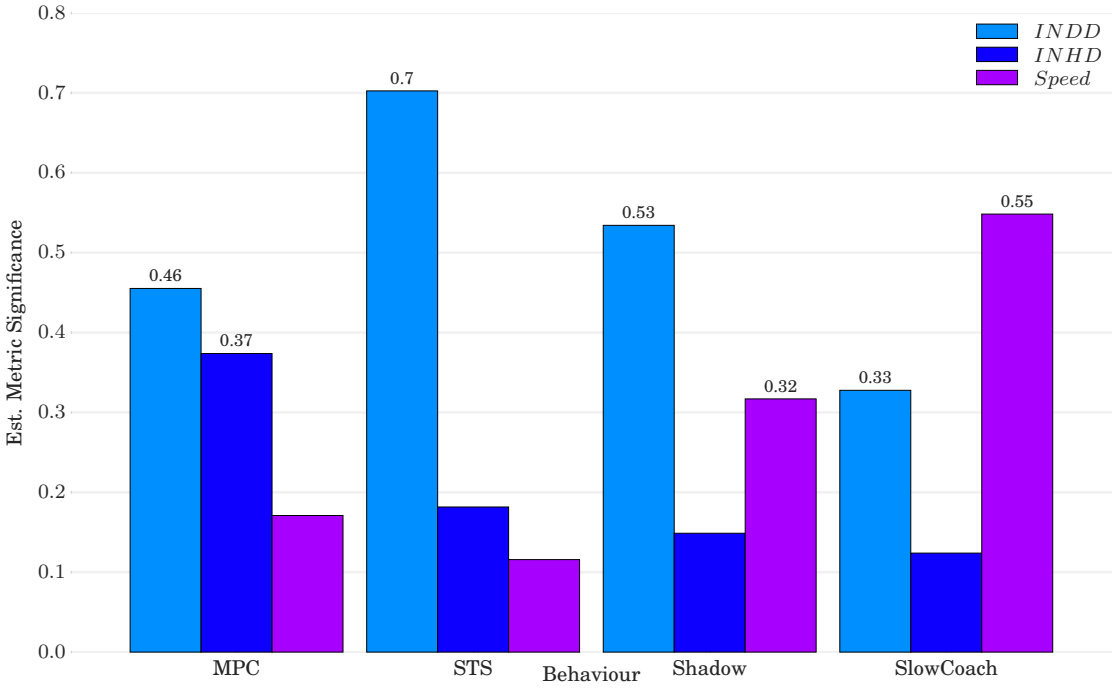


Figure 6.3: Physical Metric Features (X_{phys})

weight permutations required to come to a conclusion about the behaviour under observation.

However, these approximated signatures have no information regarding the "sign" of the g,b comparison vectors from (**??**), i.e. there is no hint as to whether the relationship is $g(x) \mapsto \max, b(x) \mapsto \min$ or $g(x) \mapsto \min, b(x) \mapsto \max$)

One option would be to go back to the regression point and expand the combination options

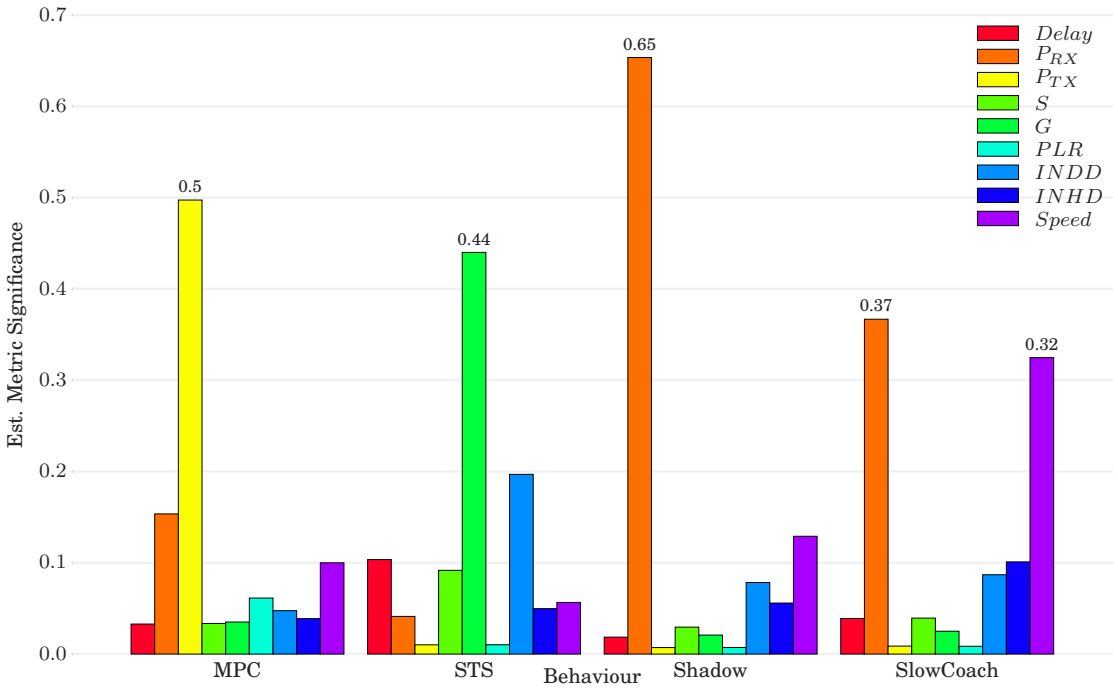


Figure 6.4: Multi Domain Metric Features (X_{merge})

to include negative values, signifying inverted g,b relationships, however this is combinatorially explosive.[1] Instead, the "significance" weight is permuted against it's possible combinations of "flips", i.e. for $X_s = [0.3,0.4,0.01,0.02,0.27]$ could also be $X_s^p = [0.3,-0.4,0.01,0.02,0.27]$ and so on. This sign permutation is filtered based on a threshold value (0.01), so for all indices below that threshold will not be permuted on, halving the number of combinations required for each indices eliminated. This reduces the number of additional assessments required from 1.9×10^6 to approximately 500 (when applied to all nine metrics) for each experimental run.

The best of these permutations is selected to both maximise the (correct) deviation between each nodes trust perspectives and to minimise the trust value reported for the misbehaving nodes; $\Delta T \rightarrow \max^+$ (**??**, results summarised in **??**). Additionally, a "False Positive" assessment, ΔT^- (**??** shown in **??**) which encapsulates the average false positive selection rate.

$$\Delta T_{ix} = \frac{\sum_{j\neq x}\left(\overline{T_{i,j}}^{\forall t}\right)}{N-1} - \overline{T_{i,x}}^{\forall t} \tag{6.7}$$

$$\Delta T_{ix}^{-} = \frac{\sum_{j\neq x}\Delta T_{ij}}{N-1} - \overline{\Delta T_{i,x}}^{\forall t} \tag{6.8}$$

Where i is a given observer, x is the known misbehaving node, $\overline{T_{i,j}}^{\forall t}$ is the average weighted trust assessment of node j observed by node i across time and N is the number of nodes.

[1] The current version of this analysis uses three metric weights; ignored, standard, emphasised, giving $3^9 = 19683$ combinations. Expanding this to include inverted standard and inverted emphasised weights would raise that to $5^9 = 1.9\times10^6$

Conceptually, ΔT_{ix} represents the "Relative Distrust" of the target node x, as the difference in trust value from $0 \rightarrow 1$, the higher the value the larger the "drop" in trst of the misbehaviour compared to the cohort. ΔT_{ix}^{-} represents the average ΔT_{ij} for all other nodes, representing the likelihood of another node being as highly distrusted as x, where positive values indicate that x is not the obvious outlier, negative values indicate that x is a very clear outlier, and near-zero values indicate a difficulty in selection of any outlier from the cohort.

The "best" weight permutations, as shown in **??**, are applied to untrained datasets for these results.

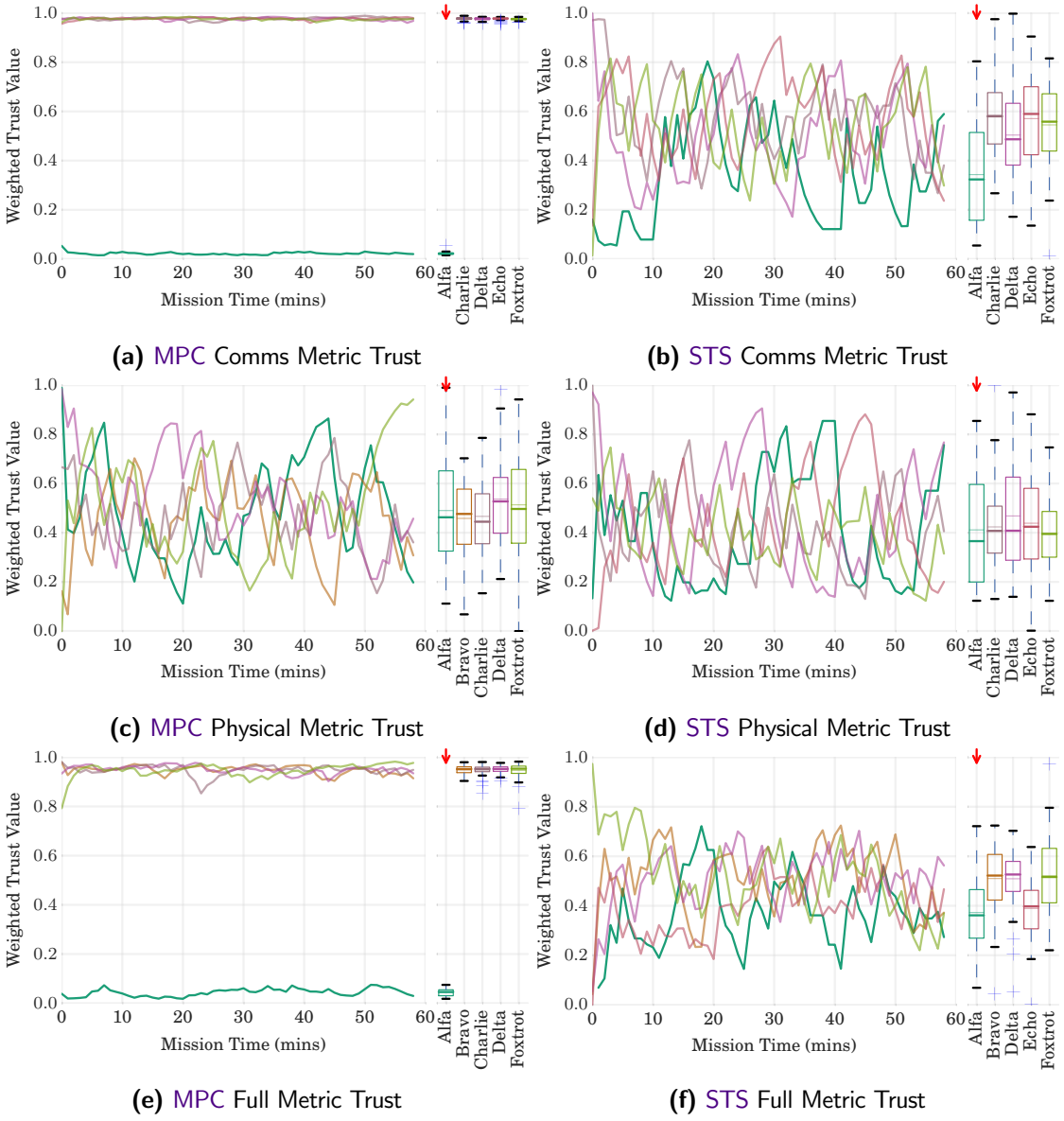
This is a departure from the Dixons Q-Test applied in **??**, as the number of metrics in use, and the recognised variability in response to different metrics makes a simple outlier assessment unfairly naive for performance assessment. The primary motivation to this work is to generate weights that (in the case of a single attacker) induce the largest Trust reduction observable by as many nodes within the network. This motivation is encapsulated in the ΔT_{ix} assessment, and allows increased abstraction so as to assess the variability of metrics in use, and to assess the usefulness of such generated "Alternate" or "Synthetic" domains.

An exemplar subset of the results is shows in Figs **??**- **??**, with the "misbehaving node" indicated above its distribution plot.

The most intuitively "Communications" behaviour, MPC, scores comfortably in the 90th percentile range in both Communications Domain (**??**) and Full Domain (**??**) trust assessments. As seen in **??**, both the "Full" and "Comms" metric optimisations heavily weigh P_{TX}, and as this is the metric directly modified by the misbehaviour, it is expected that this is easily discernible using these domain weights. However when this communications information is unavailable, as is the case in the use of Physical Domain metrics alone in **??**, the misbehaving node (Alfa) is completely indiscernible compared to the other nodes, with all nodes in the cohort tending to a trust value of 0.5. How this discernibility would fare under varying emphasis of behaviours is an open question.

Under the most "subtle" behaviour; STS, where no direct metric is being modified in operation, but where the behaviour is effectively in the "Application layer" of the networking stack, the picture is far more murky. Comparing Figs **??** and **??**, while there is a reasonable dip in the misbehavers trust assessment, the high level of variance across the cohort is such that this "mistrust" triggering is neither consistent or obvious. From **??**, the metric of import is G, the Offered Load on the network, and given it's negative weighting, this matches the expectation that the node doing "less than it's fair share" is potentially misbehaving. Unfortunately this is the case across the STS responses, where in Table **??** we have summarized out general results, STS has by far and away the lowest average ΔT in all domains. Interestingly however is the observation that Comms-only trust performs slightly better than Full trust weighting.

Referring to Figs **??** and **??**, it's clear that the offered load (G) is the almost singular feature of this behaviour, due to it's almost completely logical behaviour that is only loosely coupled to the state of the environment. The massive emphasis placed on load could only be diminished by putting it together in a larger ensemble. In Figs **??** and **??**, the misbehaving node is much more obvious than in STS, which is moderately surprising for a physically-focused behaviour. Further, there is a roughly 20% improvement when incorporating the full metric space.



(a) MPC Comms Metric Trust

(b) STS Comms Metric Trust

(c) MPC Physical Metric Trust

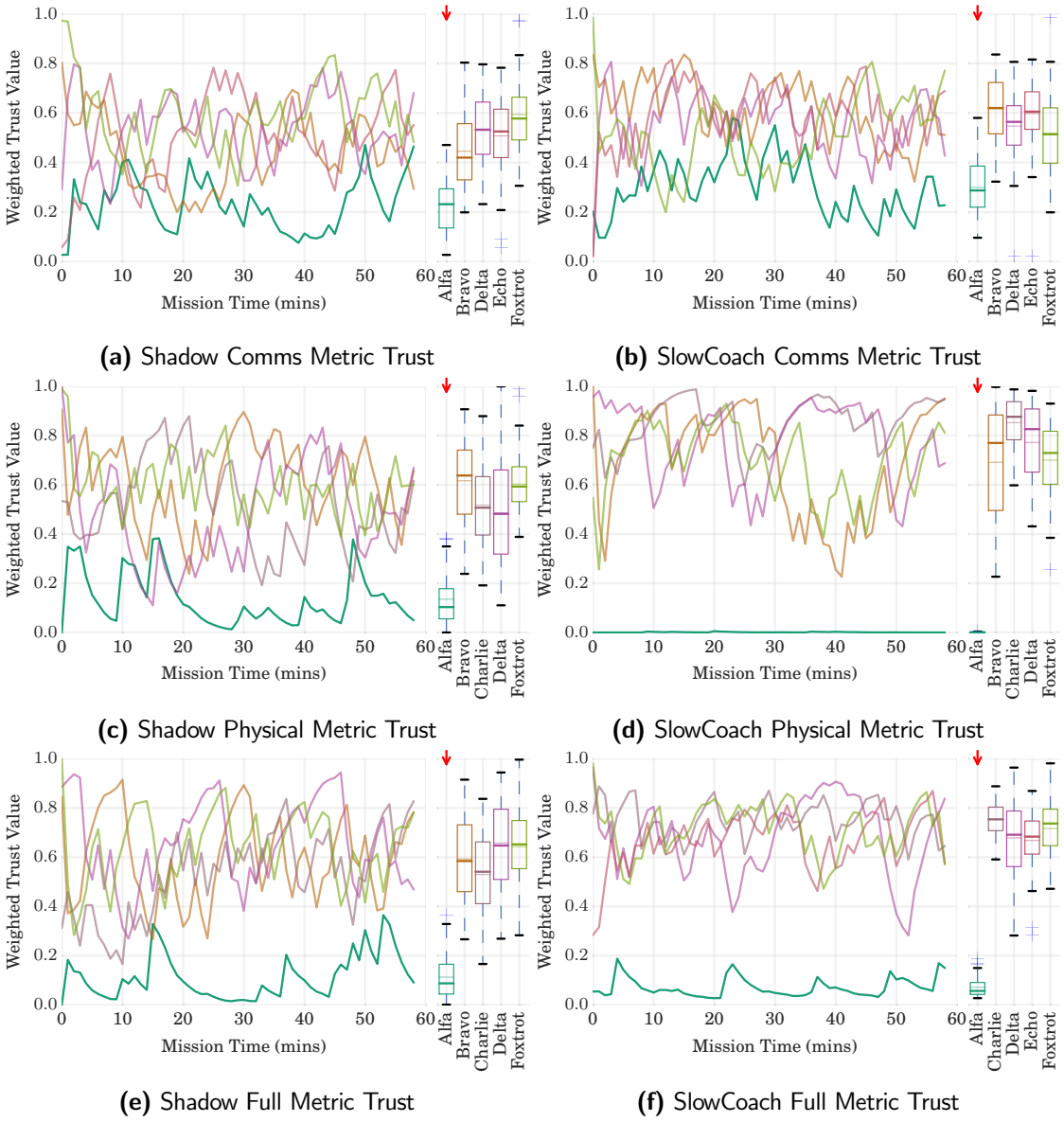
(d) STS Physical Metric Trust

(e) MPC Full Metric Trust

(f) STS Full Metric Trust

Figure 6.5: T^{MTFM} assessments for MPC and STS "Communications" behaviours

From Table **??**, the Shadow behaviour is the most consistently detectable behaviour across selected metric domains, which further suggests at the opportunity of correlating metric assessments across multiple domains.



(a) Shadow Comms Metric Trust

(b) SlowCoach Comms Metric Trust

(c) Shadow Physical Metric Trust

(d) SlowCoach Physical Metric Trust

(e) Shadow Full Metric Trust

(f) SlowCoach Full Metric Trust

Figure 6.6: T^{MTFM} assessments for Shadow and SlowCoach "Physical" behaviours

Table 6.2: ΔT across domains and "proposed" behaviours targeting known misbehaving node

Domain \ Behaviour	MPC	STS	Shadow	SlowCoach	Avg.
Full	0.81	-0.03	0.42	0.60	0.45
Comms	0.85	0.04	0.19	0.26	0.34
Phys	0.04	0.00	0.39	0.69	0.28
Comms alt.	0.85	0.03	0.38	0.45	0.43
Phys alt.	0.48	0.03	0.42	0.63	0.39

Table 6.3: ΔT^{-} False Positive assessments across domains and "proposed" behaviours across non-misbehaving nodes

Domain \ Behaviour	MPC	STS	Shadow	SlowCoach	Avg.
Full	-0.16	0.01	-0.08	-0.12	-0.09
Comms	-0.17	-0.01	-0.04	-0.05	-0.07
Phys	-0.01	-0.00	-0.08	-0.14	-0.06
Comms alt.	-0.17	-0.01	-0.08	-0.09	-0.09
Phys alt.	-0.10	-0.01	-0.08	-0.13	-0.08

Table 6.4: Optimised metric vector weights per domain trained upon and behaviour targeted

Domain, Behaviour \ Metric		*Delay*	P_{RX}	P_{TX}	S	G	PLR	$INDD$	$INHD$	*Speed*
Full	MPC	-0.033	0.154	0.495	0.034	-0.035	0.062	-0.047	-0.039	-0.101
	STS	-0.106	0.042	0.010	0.095	0.438	0.010	-0.194	-0.049	-0.055
	Shadow	0.019	0.656	0.007	-0.030	-0.021	0.007	-0.081	-0.054	-0.125
	SlowCoach	0.040	0.373	0.009	-0.042	-0.025	0.009	-0.087	0.099	-0.316
Comms	MPC	0.045	0.068	0.665	0.029	-0.043	0.150			
	STS	0.098	0.083	0.047	0.118	-0.608	0.046			
	Shadow	-0.358	0.279	0.025	0.119	0.193	0.024			
	SlowCoach	-0.082	0.309	0.021	0.090	0.478	0.020			
Phys	MPC							-0.439	-0.383	-0.178
	STS							-0.729	-0.164	-0.108
	Shadow							-0.555	-0.142	-0.304
	SlowCoach							-0.285	-0.118	-0.597
Comms alt.	MPC			0.731	0.019	-0.024	0.211	-0.014		
	STS			0.040	-0.131	-0.444	0.038	-0.348		
	Shadow			0.033	-0.124	-0.104	0.032	-0.707		
	SlowCoach			0.029	-0.164	-0.184	0.028	-0.595		
Phys alt.	MPC	0.043	0.389					-0.311	-0.075	-0.183
	STS	-0.356	0.095					-0.235	-0.135	-0.179
	Shadow	0.081	0.577					-0.097	0.070	-0.175
	SlowCoach	-0.106	0.309					-0.067	0.099	-0.420

6.2 Metric Subset Analysis

So far, the ability to generate and test the "best" metric weighting schemes across domains has been demonstrated, optimising for the highest levels of selectivity between fair and expected misbehaviours. However, these metric weighting schemes have been constructed into "domains" based on natural experience in the operating environment. Through observation of the metric significances shown in **??**, potential alternative "domains" were generated through simple observation of the visual result. In order to remove the human element, and investigate the wider impact and potential optimisation of this "optimum subdomain" subset idea, the idea whether these intrinsic domains (Physical and Communications) can be improved upon by removing the assumption that "Communications" behaviours are best identified through the use of Communications metrics is tested.

To accomplish this, the discussed analysis from metric weight significance regression and generation to $\Delta T_{ix}/\Delta T_{ix}^{-}$ validation is performed for all combinations of the $M=9$ explored metrics with three or more metrics ($k \geq 3$).

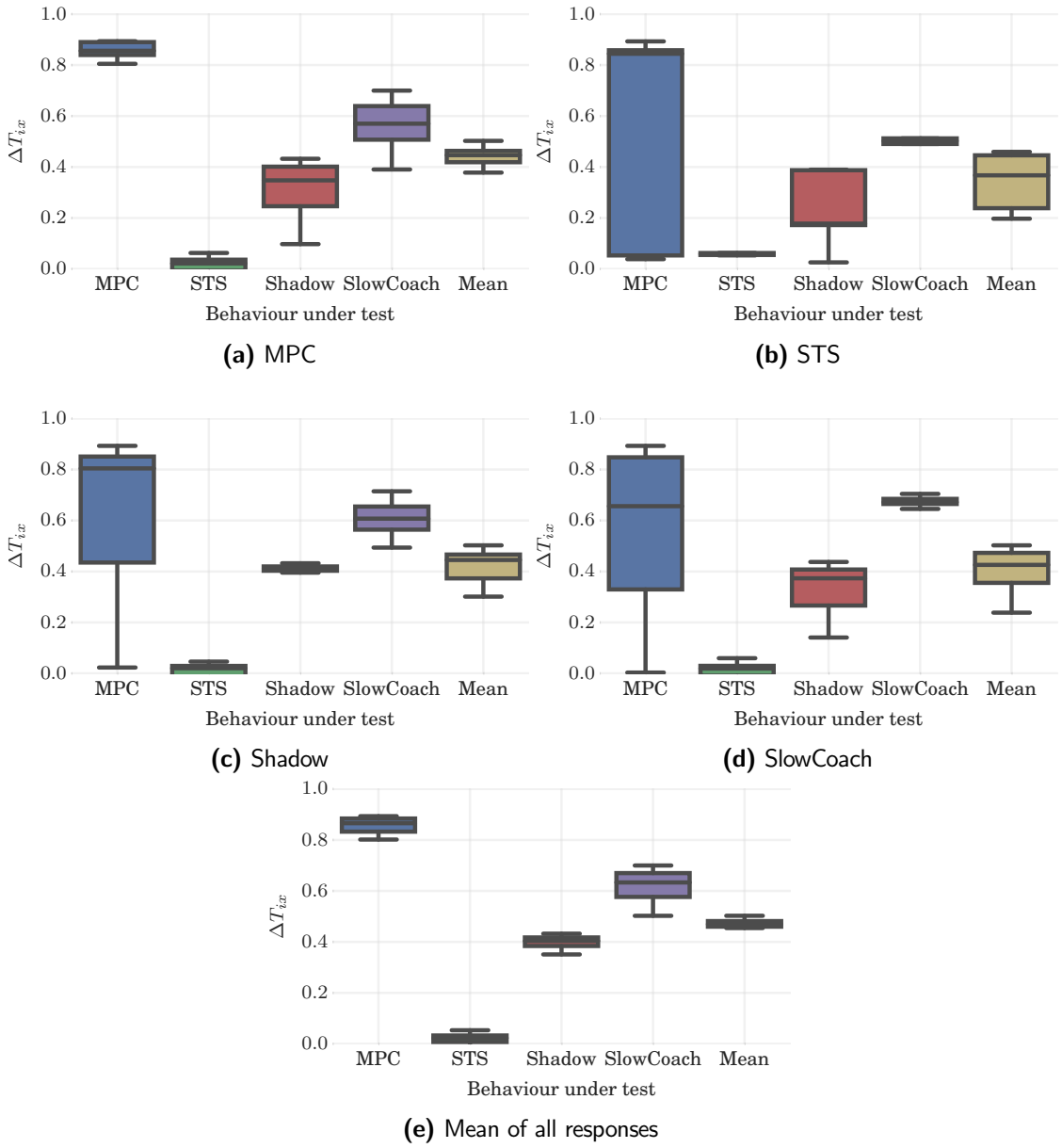
From this brute-force approach [2], a small investigation can be made in to both the performance of metric subsets, and the potential redundancies between metrics. For instance, one could expect that P_{RX} would be almost always directly related to the expected positioning between nodes and therefore to INDD. However, a counter hypothesis would be that this redundancy is present in the "Fair" case, but in misbehaving cases, the discrepancy between P_{RX} and INDD could indicate or characterise a particular misbehaviour.

The True Positive (ΔT_{ix}) is again used to indicate the overall performance of a particular synthetic domain (i.e. a synthetic domain created by the arbitrary selection of a given set of trust metrics for optimisation and assessment). **??** shows the distribution in ΔT_{ix} for each behaviour for the top 10% of this simple mean. As has been shown before, it's immediately clear that MPC and SlowCoach are the more responsive behaviours to detect across the metric space, with Shadow being slightly more difficult and STS remaining as challenging to detect as in earlier attempts. This is disappointing, as it was hoped that *some* combination of metrics in relative isolation would be capable of highlighting this behaviour in a more convincing manner, but it is clear that the Application level nature of the STS attack is avoiding significant impacts across all the metrics currently applied.

FUT: It'd be good to apply this to "blind" behaviours, but that's a very complicated issue in terms of anomaly detection etc. Should work though as long as it's not Application-specific like STS

Across the figures in **??**, when behaviours are targeted for this meta-optimisation, naive as it is, the variability in response performance in other, unoptimised, misbehaviours is greatly increased, indicating that, at least in the case of attempting to *identify* misbehaviour, that this is a case of over-optimisation, and that in the majority of cases, optimising for the mean response (**??**) is sufficient to get a strong and consistent performance across misbehaviours.

[2] each "run" consisting of $\binom{M}{k}$ individual weight assessments, with the same multiplicity of 8 experimental runs per scenario per target node as before in previous runs



(a) MPC **(b)** STS

(c) Shadow **(d)** SlowCoach

(e) Mean of all responses

Figure 6.7: Variability in cross-behaviour performance of Top 10% of synthetic domains based on individual misbehaviours and the overall mean response

This is shown explicitly in **??**, where the selective performance ratio between a targeted subset optimisation and the "mean" subset optimisation demonstrates that in all cases but STS, targeting specific behaviours for this domain synthesis does not meaningfully improve the true-positive behaviour of trust assessment, and in general decreases the detection accuracy. The abstractions for the generation of **??** are shown in **??**.

Looking at the performance of these targeted synthetic domains with respect to the metrics used, **??** shows the metric selection correlations across metrics and behaviours. From this it is observed that while there is a relatively consistent relationship between "Physical" metrics (INDD, INHD and Speed) and "Physical" misbehaviours (SlowCoach and Shadow), this is far from consistent across the domain space.

Table 6.5: Top 5 performing synthetic domains, targeting MPC, and including their performance in detecting other misbehaviours

Behaviour ΔT_{ix}					Metrics in Synthetic Domain								
MPC	STS	Shadow	SlowCoach	Mean	*Delay*	P_{RX}	P_{TX}	*S*	*G*	*PLR*	*INDD*	*INHD*	*Speed*
0.89	0.01	0.35	0.54	0.45	✓	✓	✓					✓	
0.89	-0.03	0.17	0.64	0.42	✓		✓		✓			✓	✓
0.89	0.05	0.12	0.46	0.38	✓		✓	✓				✓	
0.89	0.04	0.35	0.55	0.46	✓	✓	✓	✓				✓	
0.89	-0.03	0.27	0.49	0.41			✓	✓			✓	✓	

Table 6.6: As in ??, but targeting STS

Behaviour ΔT_{ix}					Metrics in Synthetic Domain								
MPC	STS	Shadow	SlowCoach	Mean	*Delay*	P_{RX}	P_{TX}	*S*	*G*	*PLR*	*INDD*	*INHD*	*Speed*
0.86	0.06	0.37	0.49	0.45	✓		✓	✓		✓	✓		
0.84	0.06	0.39	0.51	0.45	✓		✓			✓	✓		
0.83	0.06	0.03	0.02	0.23	✓		✓	✓		✓			
0.04	0.06	0.18	0.68	0.24	✓				✓	✓			✓
0.05	0.06	0.17	0.51	0.20	✓					✓		✓	✓

Table 6.7: As in ??, but targeting Shadow

Behaviour ΔT_{ix}					Metrics in Synthetic Domain								
MPC	STS	Shadow	SlowCoach	Mean	*Delay*	P_{RX}	P_{TX}	*S*	*G*	*PLR*	*INDD*	*INHD*	*Speed*
0.49	-0.00	0.44	0.66	0.40		✓					✓	✓	✓
0.81	-0.03	0.43	0.68	0.47		✓	✓			✓	✓		✓
0.81	0.02	0.43	0.66	0.48	✓	✓	✓			✓	✓		✓
0.78	-0.02	0.43	0.62	0.45		✓	✓			✓	✓	✓	✓
0.40	-0.01	0.43	0.63	0.36		✓				✓	✓	✓	✓

Looking back at the performance of the previously used Native and Alternative Domains, **??** can be extended to include results from the "Best" synthetic domains for each misbehaviour, shown in **??**. As stated, it is not surprising that these Behaviour-optimised synthetic domains perform better at maximising ΔT_{ix} in their targeted behaviours, with an incurred reduction in the average response to other misbehaviours.

In terms of a comparison between the previously generated "Alternate" domains, which were made by visual inspection of the returned relevance from **??**, the "SlowCoach" synthetic domain uses almost all the same domains as the "Phys Alt." domain, such that the synthetic domain leaves the Delay metric out, and in all cases except for STS, outperforms the Alternate domain.

Table 6.8: As in ??, but targeting SlowCoach

Behaviour ΔT_{ix}					Metrics in Synthetic Domain								
MPC	STS	Shadow	SlowCoach	Mean	*Delay*	P_{RX}	P_{TX}	S	G	PLR	$INDD$	$INHD$	*Speed*
0.47	0.00	0.37	0.72	0.39	✓	✓		✓					✓
0.67	-0.03	0.39	0.72	0.44	✓	✓			✓				✓
0.52	0.02	0.42	0.71	0.42		✓			✓		✓		✓
0.37	0.03	0.40	0.71	0.38	✓	✓					✓		✓
0.33	-0.02	0.40	0.71	0.36		✓		✓			✓		✓

Table 6.9: As in ??, but targeting the mean response across misbehaviours

Behaviour ΔT_{ix}					Metrics in Synthetic Domain								
MPC	STS	Shadow	SlowCoach	Mean	*Delay*	P_{RX}	P_{TX}	S	G	PLR	$INDD$	$INHD$	*Speed*
0.88	0.03	0.42	0.69	0.50		✓	✓		✓		✓		✓
0.87	0.03	0.42	0.68	0.50	✓	✓	✓		✓		✓		✓
0.89	0.04	0.37	0.69	0.50		✓	✓		✓				✓
0.87	0.02	0.42	0.67	0.50	✓	✓	✓				✓		✓
0.88	0.04	0.38	0.68	0.49		✓	✓	✓	✓				✓

Table 6.10: Averaged summary of top 10% for each targeted behaviour, with the average ratio of occurrence of each metrics for the summarised synthetic domains

Targeted Behaviour	Behaviour ΔT_{ix}					Metrics in Synthetic Domain								
	MPC	STS	Shadow	SlowCoach	Mean	*Delay*	P_{RX}	P_{TX}	S	G	PLR	$INDD$	$INHD$	*Speed*
MPC	0.89	0.02	0.25	0.53	0.42	0.65	0.40	1.00	0.50	0.40	0.00	0.20	0.70	0.45
STS	0.58	0.05	0.25	0.43	0.33	0.80	0.40	0.55	0.70	0.30	0.65	0.30	0.35	0.35
Shadow	0.70	0.00	0.43	0.63	0.44	0.50	1.00	0.65	0.05	0.45	0.75	1.00	0.70	0.95
SlowCoach	0.52	0.00	0.38	0.70	0.40	0.50	0.80	0.25	0.35	0.65	0.15	0.65	0.05	1.00
Mean	0.87	0.02	0.40	0.67	0.49	0.55	1.00	1.00	0.35	0.70	0.20	0.60	0.20	1.00

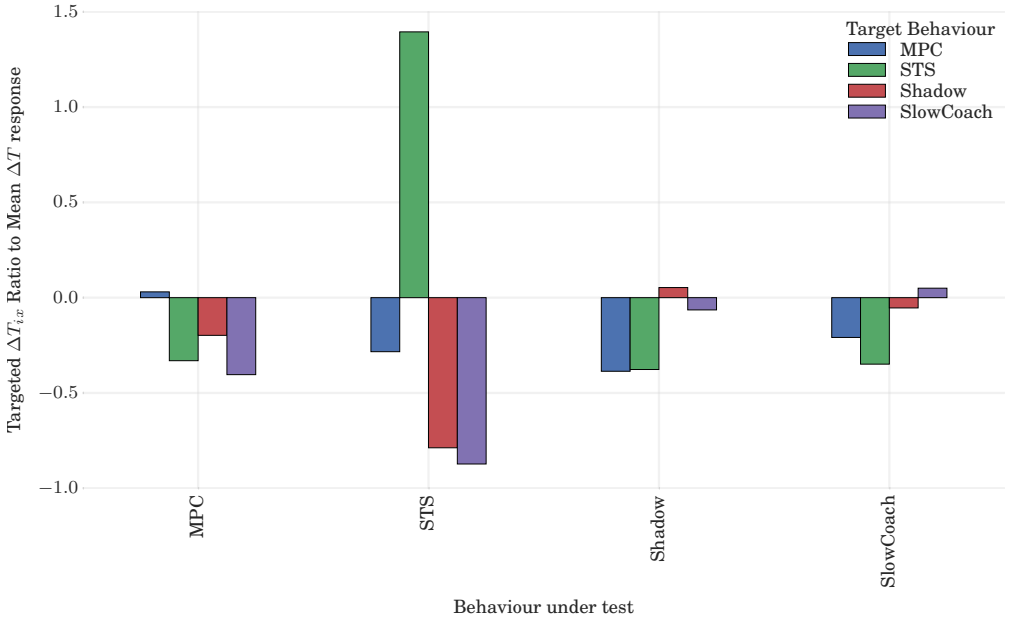


Figure 6.8: Mean of all responses

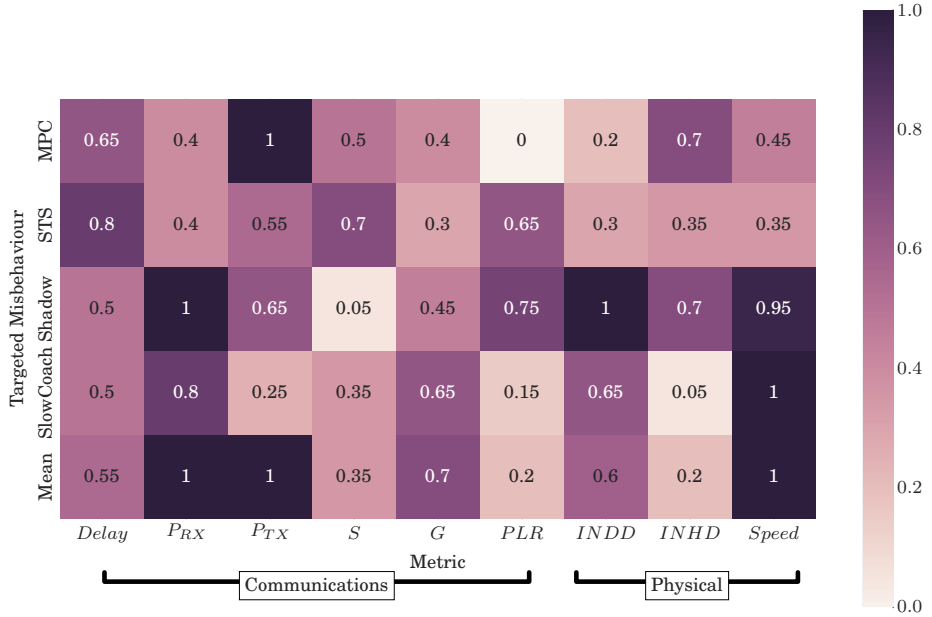


Figure 6.9: Correlations between highest performing synthetic domain metrics with respect to Targeted misbehaviours

Table 6.11: ΔT_{ix} behaviour detection performance across basic, alternate, and targeted-synthetic domains, showing the respective constituent metrics

Domain		Behaviour ΔT_{ix}					Metrics in Domain								
		MPC	STS	Shadow	SlowCoach	Mean	*Delay*	P_{RX}	P_{TX}	S	G	*PLR*	*INDD*	*INHD*	*Speed*
Basic	Full	0.81	-0.03	0.42	0.60	0.45	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Comms	0.85	0.04	0.19	0.26	0.34	✓	✓	✓	✓	✓	✓			
	Phys	0.04	0.00	0.39	0.69	0.28							✓	✓	✓
Alternate	Comms alt.	0.85	0.03	0.38	0.45	0.43				✓	✓	✓	✓		
	Phys alt.	0.48	0.03	0.42	0.63	0.39	✓	✓					✓	✓	✓
Synthetic	MPC	0.89	0.01	0.35	0.54	0.45	✓	✓	✓					✓	
	STS	0.86	0.06	0.37	0.49	0.45	✓		✓	✓		✓	✓		
	Shadow	0.49	-0.00	0.44	0.66	0.40		✓					✓	✓	✓
	SlowCoach	0.47	0.00	0.37	0.72	0.39	✓	✓		✓					✓
	Mean	0.88	0.03	0.42	0.69	0.50		✓	✓		✓		✓		✓

6.3 Selectivity Performance of Optimised Domains

Having arrived at a set of Basic, Alternate, and now Synthetic domains, optimised for maximising the induced "drop" in trust, (i.e. maximising ΔT_{ix}), an estimate of the actual performance of these assessments must be made. This is accomplished using an extension of the Dixon classifier used in **??**, such that rather than detecting outliers based on a physical metric, outliers are detected based on the differential trust value given by a given optimised weight.

This is initially tested in what can be considered the "best case"; using the previously optimised domains above to weight a collection of over 5000 execution runs in each of the four domains, and using this multi-domain trust vector in Dixons' Q-Test. (Again, Q^{95})

The results from this detection test are shown in **??**. Comparing the synthetic domain results to the existing alternate and basic domains, in all cases except for SlowCoach, the targeted domain performs marginally better than any other domain, but the margin in this is very slim.

STS continues to evade direct identification despite having optimistic metric and domain significances identified; across all metric weightings and domains applied, the 7% detection rate arrived at with its synthetic domain (matching it's false positive rate) may be the best that can be expected from this methodology. This is particularly disappointing compared to the "manual" classifier developed in **??**.

you were here when you went to bed

T

target_var target_domain	++ MPC	STS	Shadow	SlowCoach	-+ MPC	STS	Shadow	SlowCoach
Full	1.00	0.03	0.63	0.98	0.00	0.07	0.00	0.0
Comms	1.00	0.05	0.18	0.39	0.00	0.03	0.04	0.0
Phys	0.03	0.02	0.40	0.85	0.12	0.09	0.00	0.0
Comms alt.	1.00	0.03	0.22	0.43	0.00	0.15	0.02	0.0
Phys alt.	0.54	0.04	0.62	0.97	0.00	0.06	0.00	0.0
MPC	1.00	0.04	0.62	0.80	0.00	0.04	0.00	0.0
STS	1.00	0.07	0.24	0.45	0.00	0.07	0.00	0.0
Shadow	0.57	0.02	0.64	0.95	0.00	0.07	0.00	0.0
SlowCoach	0.70	0.04	0.71	0.88	0.00	0.06	0.00	0.0
Mean	1.00	0.02	0.70	0.93	0.00	0.12	0.00	0.0

Table 6.12: Selectivity performance using Domain-Trained weight vectors using a Dixons Q based limit-classifier

6.4 Conclusion

In this chapter we demonstrate that in harsh environments, multi-domain trust assessment can perform better on average than single-domain counterparts, both in terms of robustness and sensitivity, but also covering a wider region of the potential behaviour space,

The extension of the methodologies of multi-vector trust into the marine space are already demonstrated, however including information from physical observations of actors in a network enables the detection and identification of a much wider range of behaviours. We also demonstrate a method for assessing trust metrics in harsh environments in terms of their relative significance, and a method for establishing classification signatures for misbehaviours. Finally, the synthetic generation of abstract metric domains is explored, where it is found that in most cases, optimising for generalised performance prevents response-overfitting, and provide the strongest deviations in observed Trust when targeted specifically, further supporting the use of a mixed-domain approach for Trust across domains for identification of misbehaviours.

It is to be noted that this presented method is significantly more computationally intensive than the relatively simple Hermes / OTMF algorithms communications only algorithms, and is exponential in complexity as metrics and/or domains are added. The repeated metric re-weighting required for real time behaviour detection is therefore an area that requires optimization. More work needs to be done to characterise how worthwhile this approach is compared to a separate synthesis approach where by MTFM-style trust is generated and assessed on a per-domain basis and subsequently fused.

For greater fidelity and more optimal results, a wider range of weights can be used in the initial regression step; however this is computationally expensive given that weighting is applied to

each perspective (i.e. observer/target node pair) for each trust assessment time step, presenting 15 perspectives at each time interval in the 6 node case.

Every effort has been made to avoid over-training the dataset, using cross validating sampling for regression and "best weight" generation, however more meta-analysis is required to further demonstrate the optimality of this process.

Chapter 7

Conclusions & Future Work

7.1 Conclusions

The use of MANET architectures in the UAN space requires a fundamental reassessment of the security and reliability mechanisms and performance of such networks in the challenging underwater environment. The strengths of MANET architectures inherently produce decentralised, self-organised, collaborative networks that strive towards efficiency and performance where all network members perform fairly. However, with the increasing introduction of autonomy into the general MANET space, this assessment of "fairness" is simultaneously an assessment of the capability of the network, and of the "trustworthiness" of the autonomous nodes within it.

The original MANET architecture was designed with no in-built defence or security capabilities, and as such, threat mitigation mechanisms been superimposed over time to protect against fundamental vulnerabilities in the architecture due to assumptions of fairness, the use of open, wireless links with "fuzzy" operational boundaries, highly mobile nodes inducing dynamic topologies, and constrained power/computation/locomotion/communications resources. These mechanisms vary, from evidence based cryptographic security such as centralised trusted third parties, a-priori shared secrets or one-time-pads, to fully decentralised PKI systems. However, these classical security measures require significant investments in memory and computational power; communications channel occupancy; and inherently rely on relatively short delays between links and from end-to-end points in the network.

In the terrestrial realm, the increasing computing power of devices such as mobile phones has enabled the creation of pervasive, end to end security. However, as discussed in **??**, these assumptions of channel availability and low-latency do not hold in the underwater acoustic space, which is massively variable in channel capacity and delay-response. As such, regardless of on-board computing power, alternative, decentralised methods for ensuring the integrity of the network and its operations is essential for expanding the applications of MANET architectures in this space.

These applications vary from defensive patrolling, ASW and MCM, to pervasive environmental monitoring, and in almost all cases, current military and commercial implementations benefit from leveraging individual node autonomy in a distributed architecture to bring down development and operating costs, increase system efficiency, and fundamentally, save humans time and in extreme cases, lives.

Trust is one alternative approach evidence based security to maintaining network integrity in the face of selfish, malicious or faulty misbehaviours in MANETs, and this approach has been well explored in the terrestrial realm. In **??**, the fundamental concept of Trust was explored, and a range of psychological, phenomenological, and technical approaches to Trust assessment and collaborative Trust were investigated. While this included excursions into the concepts of Design Trust (**??**) and the impacts of human factors on the expected performance of trusting systems (**??**), this discussion was directed towards the application of Trust to autonomous MANET, as well as currently developed methods for establishing and maintaining trust in MANETs such as the Hermes and OTMF single-metric assessment frameworks (**??**), and MTFM, which broke new ground in Trust establishment by looking at many available communications related metrics as an ensemble(**??**), and applies Grey Relational Grading and whitenization (**??**) to take assessments across the communications domain and across the network topology to assess the trustworthiness of nodes within a MANET in a distributed fashion without requiring environmental or application specific "training".

In general, Trust is "the level of confidence one agent has in another to perform a given action on request or in a certain context" (**??**), and has previously been exclusively concerned with the communications operations of networks, generally relying on measures of packet routing success to infer expectations of future packet delivery probabilities, improving route generation and mitigating threats from selfish or malicious interference (**??**). However, the UAN application area and its challenging operation and communications constraints puts unique challenges on previous assumptions about the operation of abstract trust frameworks in MANETs, and as such required reassessment.

Further, the UAN application area also highlights more general challenges to Trust in autonomous MANET; with the imposition of a highly constrained communications channel, it is difficult to maintain sufficient information about the operation of nodes in the network with high enough regularity to reliably disseminate that information across the network. Also, the high constrained physical dynamics and resource intensive communications and locomotion in UAN MANETs greatly expands the potential threat surface, particularly for DoS-style resource manipulation attacks; when locomotion is energetically expensive, if a node can selfishly get a "free ride" by minimising its mobility rather than fairly distributing effort across the network, operational mission times and overall efficiency can be impaired.

As such there is an open opportunity to explore the application of Trust methodologies to the physical domain instead of and as well as the communications domain, making such a TMFs able to identify a much greater range of potential misbehaviours and maintain both integrity and efficiency.

7.2 Contributions and Findings

Within this context, **??** initially explores the scaling differences in node distance and communications rates using a simulated agent based environmental and communications model, identifying network saturation rates using a range of mobility, scaling, and offered loads, maximising network performance in terms of a throughput-delay product. Existing TMFs methods are applied

to this established range, demonstrating that these (Hermes, OTMF and MTFM) existing frameworks are not directly suitable to the sparse, noisy, and dynamic underwater environment. While there is little that can be done to augment Hermes and OTMF in this environment, the weighted-metric nature of MTFM allows the metric space to be explored for "better" weighting vectors to detect and identify misbehaviours that are hidden in the unweighted assessment.

Having established the operation of Trust in the marine MANET environment, **??** demonstrates through statistical measures of node distribution and velocities, that malicious and faulty misbehaviours can be both detected and identified to a high selectivity using a simple tree-based classifier, using a collaborative Port Protection scenario as a targeted waypoint mobility baseline.

??, the combination of these domains is assessed. The relative significance of metrics in the specific identification of a given behaviour is performed through a full-sight random forest regression by comparing a significant number of simulated mission runs. These significances, while simply a statistical measure of the importance of a given metric in discriminating between two behaviour (i.e. a fair behaviour and the misbehaviour) provide actionable information to generate a weighted metric sequence targeting that behaviour. The identification and classification methodologies from both **??** and **??** are combined to generate a classifier with very beneficial performance.

7.3 Future Work

One of the fundamental difficulties in establishing trust in sparse, noisy networks is that it takes a significant amount of time to accumulate enough observational metric data to for actionable opinions. This problem is exacerbated by the asynchronous nature of those metrics; this is evident in the multi-domain discussion between communications and physical metrics where overheard location updates for a given node my arrive out-of-sync from other useful information about that particular relationship, but is also the case in the pure communications domain.

If these metrics are not synchronised, for instance if they are interrupt driven such as communications-based observations, generating more abstract measurements requires inherent assumptions about "how to accumulate the data while you wait". For instance, **??** and [**?**] demonstrated a periodic trust assessment framework for autonomous marine environments, in such an environment, to establish useful, generalised, data, it was necessary to wait for a relatively long time to accumulate enough data to make assessments. However, this leads to data being left in-buffer for a time before being used to make decisions, and by the time the data was collated and processed, it could be wildly different from the reality. Further, while some periods could be extremely sparse or even empty, others could be extremely busy with many records having to be averaged down to provide a 'single period' response. One solution to this would be to move from a stepping-window of trust observations as used in this work to a continuous trust log, updated on packet reception rather than waiting regular periods for packets to be analysed. Therefore, the implementation of a suitable grey sequence buffer version of the framework would be beneficial.