I.

A reputation system must ensure that the data it represents is true and has not been manipulated in any malicious manner. Unfortunately, like many systems, they are susceptible to attacks and these have two sources but take multiple forms.

- **Passive or active attacks** [1]. A passive attack is when an attacker manages to gain access to a system but does not actively alter any information, they monitor it and perform an analysis of the data or traffic. An active attack is when modification occurs.
- **Insider or outsider attacks** [1]. An insider attack is performed by an authorised party and then performs some kind of attack on the system. In contrast, an outsider attack is one whereby the attacker gains access to the system and then performs an insider attack.

There can be a wide range of sources that result in fake reviews being published and dependant on the attack type and target they fall in to one of five categories:

- **Self-promoting**. An attacker attempts to increase the reputation of their own page. An example of this kind of attack would be a user creating multiple accounts and leaving positive reviews on their business.
- **Sybil attack**. The success of an attack can be dependant on the systems ability to limit the number of accounts that can be created by one source; known as a Sybil attack [2]. If a system does not prevent an address from creating a high number of accounts or rate-limit this number then an attacker can perform a *ballot stuffing* where a number of entities leave positive reviews for the same user, or *bad mouthing*, the contrary [3][4].
- **Whitewashing**. An attacker can use a system vulnerability in order to increase their reputation or re-enter the system with a new identify, thus escaping reputational penalties [5]. An example of this kind of attack would be a user recreating their business page following a series of bad reviews of their business.
- **Slandering**. Where an attacker lowers the reputation of another user purposely. An example of this kind of attack would be a user creating multiple accounts and leaving bad reviews on a competitors business.
- **Orchestrated**. Where an attacker uses a combination of the above attacks.
- **Denial of service (DoS)**. DoS attacks can be used in order to prevent parts of the reputation system from working as they should, such as temporarily disabling the calculation of values. While availability attacks are not unique to this field they do present issues when used against disseminators as they can prevent the propagation of information throughout the system [1].

While a vast number of attack types exist, some of the most commonly seen attacks are bad mouthing and ballot stuffing [6]. As suggested by Dellarocas [4], two discriminatory seller behaviours can be introduced, negative, where a select few buyers are negative discriminated against, and positive, where a select few buyers are treated exceptionally. These four misbehaviours create a dispersion of ratings for a given user [4] and bring to light the question of the authenticity of a transaction alongside identifying discrimination in a review. Websites such as eBay and Amazon require users to have purchased a product before a review can be left, this differs to offline services where a user will seek out a review first, use the service offline and then leave a review. This creates an indirect disconnect between the reputation system and the service that is being reviewed. The lack of a certified method of validating the authenticity of a review opens the platform up to the aforementioned attack types.

Denial of Service attacks or any kind of availability attacks in relation to trust and reputation systems have not seen enough research yet to form any correlation between the two [1]. However, enterprise systems do face a real threat of DDoS attacks and the results of them can be catastrophic and as such pose a high risk to businesses. Section A.13.1 (network security management) of ISO 27001 places a constraint on the business that firewalls and network intrusion detection systems are in place to attempt to prevent them. In addition to this, any other kind of malicious attempt in gaining control over the system must be prevented as an intruder could manually change the reviews in a database in a self-promoting, whitewashing or slandering attack. In addition to this, section A.12.1.3 (capacity management) through 12.4.1 (event logging) place constraints on how GENRE will identify attacks at an earlier stage and log these. While measures exist to circumvent the aforementioned attack types, they are largely not attempts to gain control over a system and manually change the data or launch a DDoS attack but this is not to allow for the risk of a network attack to be forgotten. The ISO requirements exist and to protect both the business and customers protections must be in place. If a user does manage to gain control over the system, security practices must be in place for once control has been regained and any backups that must be restored and ISO 27001 A.10.5 (Back-up) defines an objective to maintain the integrity of the data within the system. The vast majority of risks are ones that happen over time and manipulate the systems that are in place. The risks of an attack can affect the GENRE business itself and the customers that use it; who may see a decrease in footfall to their business and this can affect their livelihood. The consequences are very real, websites like TripAdvisor are under constant threat [7][8] and just last year a man was failed for selling fake TripAdvisor reviews [9].

II.

Websites that do not operate through the use of user accounts make it more difficult to restrict the effectiveness of Sybil attacks, however, strategies to exist but their effectiveness is lower. A number of fingerprinting techniques exist in order to identify a browser, however, these can result in a number of false positives. A more effective approach in circumventing a Sybil attack would be to limit the number of reviews that a user can generate. This may be frustrating for legitimate reviewers, if an authentication and user principal system is not implemented then this is one of the few methods that will help avoid false reviews. Techniques such as I.P blocking can be effective here but they can be circumvented through the use of proxies, VPNs and the Tor network. This defence strategy can be followed by the notion of placing a trust level on a node in proportion to its age. Thereby introducing a new metric that can be taken in to consideration when calculating the overall trustworthiness of a vendor. Yu et al. [10] introduced the notion of creating a graph which identities take the form of a node and their edges are trust levels. A high level of malicious users would offset the graph and their removal would disconnect a large number of nodes.

Providing users with the ability to rate other user's reviews is another method that can help to lower the proportionality that a review is taken in to consideration when calculating the overall rating of a node. This, however, further opens the system up to a new set of vulnerabilities. In developing a number of metrics that can be used to measure the quality of a given review, the overall reputation for a node in the reputation system can be calculated accordingly. Given the aforementioned: the age of user, quality of their review and the confidence level that the user leaves fake reviews can all be used. Incorporating more metrics in to the reputation function decreases the vulnerability of the system to ratings attacks [11]. For example, Swamynathan et al [11] discuss that if a system placed enhanced weighting on a user that left a large number of reviews in the past then the system would suffer from an "increased trust by increased volume" vulnerability, even if the transactions were verified as the user could have purchased a large number of goods at a low value.

Fake reviews in systems posted by anonymous users can also be detected through the content of the review itself. Liu et al. [12] performed research in to the contents of reviews when measured against a products description and found that the content was arranged in such a manner that it matched the description, as this is the reviewers only source of information. Reviews that present this behaviour should be given a lower weighting as they appear to be false. Liu et al. [11] also proposed creating a dataset from what is believed to be genuine reviews and using this to measure submitted reviews against in order to create an authenticity score of the review that can be used when calculating the reputation of a node.

Another method that a node can encourage genuine reviews is to only allow for a review to be created if a code is entered in to the system at the time of creation. As the business is offline and requires human-to-human interaction, a code can be given to the customer at the time that they check in to the hotel or enter the museum and this can be used to vouch for the authenticity of the author.

In recent years, click farms and other services which offer fake reviews have become increasingly popular. This can result in a large number of reviews for a node or nodes that are geographically close together originating from one or a small number of IP addresses.

Buccafurri et al. [13] propose the use of a bias of a review be applied to check how far the review deviates from the expected quality of the node. Large deviations give the reputations model for a node the ability to detect if the review is outside of the expected bound and so large quantities of reviews can be given a lower weighting. Implementing such a system requires a long term training phase in order to generate review score that will result in subsequent coherent review biases. Their model is further enhanced through the use of authentication and user accounts, however, adopting it to a non-authenticated website and using the above metrics, the following variables can be added to the formula:

- f_1: This parameter is 1 if the reviewer leaves their contact information or 0 else.
- f_2: This parameter is 1 - the percentage similarity that the review is to the node's description of the business.
- f_3: This parameter is 1 - the percentage number of reviews that this node has seen from the originating IP address.
- f_4: 1 if the review contains an image, 0 otherwise. Higher quality reviews have more effort put in to them and contain images (if the website supports them).
- f_5: The percentage of like:dislikes that this review contains.
- f_6: A percentage difference that this review's length is from the average of the node.

For brevity, the formulas that Buccafurri et al. have defined have been left out, however, the result of their bias formula is a percentage difference from the review score. False reviews that do not meet the above criteria will result in a larger bias and can be assigned a proportional value to reduce the reviews impact on the overall node.

III.

Through the addition of a user account system to the GENRE platform allows for user profiles to be built, personality profile and trends. From this, consistent reviews from users in terms of their quality and homogeneous reviews in repeat visits to venues, stronger reviews can be formed. "Most online communities lack strong authentication, allowing users to obtain multiple independent online identities" [14, p. 252], however, this approach introduces the notion of using two-factor authentication for the user's account or the use of a Facebook login system. While it is possible to create multiple Facebook accounts, reviews that do not have two-factor enabled can be given a further lower rating in contrast to having two-factor enabled. Two-factor authentication requires a user to verify their phone number with either a text message or automated phone call where they enter a code. This adds an extra step of complexity to when attempting to generate false reviews as the user has to acquire multiple sim cards when creating multiple accounts. Two-factor authentication also provides the user with added security for their account in case the account is compromised or they forget their login details. Accounts which are found to have posted fake reviews can also have the Facebook account, mobile phone number and email address banned from the system and this will further deter users from posting fake reviews.

The proposed system will build up personality profiles about users based on their average spend at venues they visit, quality of their reviews and the services that they have available on their account: two-factor authentication and Facebook authentication. With the use of user accounts, it is now possible to tackle positive discrimination. Dellarocas [4] proposes using collaborative filtering techniques to identify the *nearest neighbour set* N of *b*. N is reviewers who have previously visited a venue *v* and are the nearest neighbours of *b*. If the colluders have considered that this kind of filtering happens then N will contain both fair and unfair ratings and form two further clusters and applying a divisive clustering algorithm to separate the sets is required. Therefore, the reputation estimate is the N_i cluster. Dellarocas [4] further proposes to avoid negative discrimination that vendors should not be able to see the identity of their buyers. While this does not directly translate to a venue reviewing system, it does keep in line with reviewers only being able to review venues that they are authorised to review; such as through the use of codes being handed out after their stay.

Chua et al. [15] concluded that there was a correlation between the length of a review and it's positivity, stating that negative reviews were frequently met with long reviews and ones which had short reviews were deemed to be questionable. Swamynathan et al. [14] discuss the notion of looking at reviews in both an application-wide context and a limited scope to the node in question. Deducing that if an application-wide trust context is conceived then if one node trusts the review then this reviewer should be trusted when reviewing other nodes too. However, this then raises the question of what an acceptable trust level is and how it deviates between nodes. A node could receive a large number of false reviews and have a lower confidence level in its reviews and this is kept in the context of that node. GENRE should employ a trust level in it's reviews that considers the user itself coupled with the metrics that are used to calculate the reputation for a node.

Websites such as Tripadvisor have a rating system for their users and this includes the ability to become a 'top reviewer'. Their requirements [16] for becoming a top reviewer include some points that were mentioned in bias equation in section **i**. Users which provide high quality reviews should have an increased weighting when calculating the overall reputation for a node. This too should be the case for GENRE and it is now possible with the implementation of an authentication system. Following this, some websites, such as StackOverflow, restrict the abilities that a user has until they reach a certain reputation on the site; such as disallowing the creation of questions until the user has gained 100 points through constructive comments. GENRE too should follow this methodology and disallow the creation of reviews until the user has engaged in discussion on the website through commenting. This will prevent bots from using the platform unless they are advanced in their ability to post comments that are all unique.

Duan et al. [17] investigated creating suspicion indexes for promoting and demoting of hotels in systems through the use of machine learning and ranking variation analysis. They develop a sophisticated reputation model for nodes from a variety of data sources. Most notably, their algorithm looks at the following data sources:

- **Time Interval between Posted date and Stayed date**.
- **Rating Preference**. Is a a user's attitude towards rating provision. This is a measure of how their visit was across the various reviewable attributes, such as: cleanliness, amenities etc.
- **Turning Day**. This allows for the reviews to fluctuate over time with the lifetime of the venue, generating a maximum, minimum and mean for the day in question.
- **User account existence duration**.
- **Consistency of reviews**. This allows for the algorithm to include how the user generally views the places that they visit.

Dependant on the outcome of this function, the review may be identified as attempting to promote or demote a certain node in the leaderboards. If a review author fails to interact with the discussions that take place on their reviews a new metric can be formed on their profile that assigns a value to their interactivity with the platform.

For a set of suspicious reviews, Ranking Variation Analysis (RVA) is performed to identify if the ranking correlates with potential benefit [17] and if this is observed than the review can be flagged as suspicious. This methodology strongly benefits from other users marking reviews as helpful or promoting discussion on the platform as these are seen to be more engaging reviews and this metric can be provided to the algorithm when processing reviews. Using this functionality would further help in reducing the number of false reviews that would be found on GENRE while simultaneously providing more functionality to the end user through the use of comments and review interactions; functionality that is found on social networks. If a review is flagged as being false then a member of staff can be identified in order to assess the review and choose to remove it and potentially ban the user after a number of offences. This manual intervention is especially important during the early stages of GENRE to ensure that the system is operating as expected but even more so in the advent of generating a false dataset for implementing a machine learning model. If a user chooses to incorporate Facebook authentication to their profile, additional data can be gathered from that profile such as their friends list, their hometown and other data sources. Websites such as TripAdvisor [18] have developed the ability to detect whether or not a review left on a node is from a friend or family member of someone that works there and this is against their use policy. Collecting this data over time will allow for the machine learning system to continuously evolve over time and develop more methods in spotting false reviews left at nodes. The key here is to develop a system that becomes smarter over time, is able to correlate more data sources and reduce the risk of false reviews being left. Using this in conjunction with network intrusion systems will increase the chance of the system being fair for all to use.

In addition to this the ISO 27005 standard requires continual monitoring of risks and false reviews are an ongoing risk. Security practices are an ongoing battle between malicious users developing more advanced attacks and corporations developing new strategies to circumvent them and it is important that GENRE has live threat monitoring to ensure that this risk is managed and minimised. Notifying users that a review has been removed is also important in case the system has falsely identified a review and alerting them allows for the review to be moderated.

While the proposed system for GENRE does place a large number of constraints in to both the derived requirements and the business that is presenting itself online in the system, it is important to keep user experience in mind. It is important that the implementation of this system does not disfavour genuine user reviews just because a user does not choose to activate two-factor authentication in the system or because they choose to stay in lower quality hotels.

IV.

From the points made in the previous answer it is clear that implementing an authentication system in to the platform provides a clear benefit in the ability to identify false reviews in the GENRE website. This allows for a user profile to be built up but also simultaneously prevents users from accessing the platform until their accounts have been verified through the use of a mobile phone or through the Facebook platform. However, through the investigating methods that can be used to prevent false reviews in systems that do not allow for this, a number of interesting methods have been identified. Most notably, the use of machine learning to detect patterns between nodes and the users, IP rate limiting and the correlation between review content and node descriptions.

A hybrid approach that implements the technologies described in `iii`, machine learning and content review analysis may further enhance the ability to detect false reviews. While IP rate limiting has been mentioned, this is something that should already be implemented in a secure network through an intrusion prevention system (IPS) and the networks distributed denial-of-service (DDoS) mitigation strategy and this should be in conjunction with a predefined rate limit for that is acceptable for users to make before a HTTP 429 Too Many Requests response is returned as par the RFC 6585 specification [19]. While a DDoS attack may be a lower risk, the consequences of can be severe and result in down-time of the application and as such this is a high risk.

With so many attack methods existing in an online reputation system and the consequences being severe - a decrease in business footfall is a possibility - it is important that all potential avenues are explored and covered that could result in a false review being made. While this may seem like a large number of technologies to implement, using more primitive approaches like IP blocking and rate limiting require less resources in contrast to some of the more advanced strategies. Incorporating machine learning in to this system opens up additional avenues that can be explored to reduce the number of false reviews and advanced implementations may be able to identify false reviews using some of the existing systems and this would allow for those systems to be removed; lowering the complexity of the application.

A hybrid system results in an approach that utilises both hardware and software technologies to reduce the risk of availability attacks, users gaining control of the system, and users leaving false reviews on the website. This in turn reduces the overall risk of an attack and ensures that the GENRE platform is ISO 27001 compliant also. However, in order to be fully compliant with the standard the business must also fulfil the remaining areas of the standard with their internal business practices.

References

[1] D. Fraga, Z. Bankovic, and J. Moya, "A taxonomy of trust and reputation system attacks," *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*, 2012. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6295956

[2] M. Castro and R. V. Renesse, *Peer-to-peer systems IV: 4th international workshop, IPTPS 2005: Ithaca, NY, USA, February 24-25, 2005: revised selected papers*. Springer, 2005.

[3] H. AÌĄlvaro, *Computational intelligence in security for information systems: 2nd international workshop ; proceedings*. Springer, 2009.

[4] C. Dellarocas, "Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior," *Proceedings of the 2nd ACM conference on Electronic commerce - EC 00*, 2000.

[5] M. Feldman, C. Papadimitriou, J. Chuang, and I. Stoica, "Free-riding and whitewashing in peer-to-peer systems," *Proceedings of the ACM SIGCOMM workshop on Practice and theory of incentives in networked systems - PINS 04*, 2004. [Online]. Available: https://www.cs.tau.ac.il/~mfeldman/papers/FPCSj06.pdf

[6] F. Buccafurri, G. Lax, S. Nicolazzo, and A. Nocera, "Fortifying tripadvisor against reputation-system attacks," *World Congress on Internet Security (WorldCIS-2014)*, 2014.

[7] T. McVeigh, "Twitter campaign takes aim at fake restaurant reviews on tripadvisor," Oct 2015. [Online]. Available: https://www.theguardian.com/travel/2015/oct/24/twitter-campaign-targets-fake-tripadvisor-restaurant-reviews

[8] G. Birchall, "'one in three tripadvisor reviews are fake' with hotels and restaurants buying glowing reviews for Âč7, investigation finds," Sep 2018. [Online]. Available: https://www.thesun.co.uk/news/7321574/tripadvisor-reviews-fake-hotel-restaurant-ratings/

[9] metaTags.other.author, "Owner of company selling fake tripadvisor reviews jailed." [Online]. Available: https://www.thecaterer.com/articles/537156/owner-of-company-selling-fake-tripadvisor-reviews-jailed

[10] H. Yu, M. Kaminsky, P. B. Gibbons, and A. D. Flaxman, "Sybilguard: Defending against sybil attacks via social networks," *IEEE/ACM Transactions on Networking*, vol. 16, no. 3, p. 576âĂŞ589, 2008.

[11] S. Liu, D. Qi, and B. Chen, "Research and design on p2p based reliable reputation management system," *2008 ISECS International Colloquium on Computing, Communication, Control, and Management*, 2008.

[12] L. Liu, X. Zhao, H. Wang, W. Song, and C. Du, "Research on identification method of anonymous fake reviews in e-commerce," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 14, no. 4, p. 1510, 2016.

[13] F. Buccafurri, M. Fazzolari, G. Lax, and M. Petrocchi, *Contrasting Fake Reviews in TripAdvisor*, 2018.

[14] G. Swamynathan, B. Y. Zhao, K. C. Almeroth, and S. R. Jammalamadaka, "Towards reliable reputations for dynamic networked systems," *2008 Symposium on Reliable Distributed Systems*, 2008.

[15] A. Y. K. Chua and S. Banerjee, "Reliability of reviews on the internet: The case of tripadvisor," *Proceedings of the World Congress on Engineering and Computer Science*, vol. 1, Oct 2013.

[16] Jc, "Become a top reviewer on tripadvisor," Jun 2018. [Online]. Available: https://leavingholland.com/become-a-top-reviewer-on-tripadvisor/

[17] H. Duan and P. Yang, "Building robust reputation systems for travel-related services," *2012 Tenth Annual International Conference on Privacy, Security and Trust*, 2012.

[18] "How does tripadvisor determine whether a review is biased?" Sep 2018. [Online]. Available: https://www.tripadvisor.co.uk/TripAdvisorInsights/w3683

[19] "Additional http status codes." [Online]. Available: https://tools.ietf.org/html/rfc6585#page-3