Seminar über Algorithmen 4.02.2014

Reputationssysteme

Adam Furmańczuk *Wolfgang Mulzer, Yannik Stein*

1 Motivation

Reputationssysteme spielen in unserem alltäglichen Leben eine bedeutsame Rolle. Ob wir eine Reise über *trivago* buchen, Elektronik im Internet einkaufen oder nach Büchern suchen, bei vielen Entscheidungen ziehen wir Bewertungssysteme zu rate. Wie ein Beispiel für so ein System aussieht, welche Angriffe es dafür gibt und mit welchen Algorithmen man sich schützen kann, wird im folgenden Vortrag an ausgewählten Themen gezeigt.

2 Reputationssysteme und Spieletheorie

Das Gefangenendilemma aus dem ersten Vortrag wird gerne als Modell verwendet, um Nashgeleichgewichte von Strategien nachzuweisen. So auch für die Grim-Strategie und Reputationelle Grimstrategie im folgenden:

2.1 Grimstrategie

Die Grimstrategie beinhaltet zu kooperieren, so lange bis eine der Seiten in der Vorrunde nicht kooperiert hat. Wenn beide Seiten diese Strategie verfolgen handelt es sich bei dem Spiel um ein perfektes Nashgleichgewicht.

Beweis. Angenommen zum Zeitpunkt t=0 wird betrogen. Nach der Grim Strategie muss per Design alle verbleibenden Runden ebenfalls betrogen werden:
Durchnittlicher Gewinn beim Betrügen: $(1-\delta)(2+\delta*0+\delta^2*0+...) = 2(1-\delta)$
Durchschnittlicher Gewinn beim Kooperieren: $(1-\delta)(1+\delta*+\delta^2+...) = 1$
$\rightarrow 1 \geq 2(1-\delta)$
$\rightarrow \delta \geq \frac{1}{2}$
Gleiches gilt für ein beliebiges $t > 0$. □

2.2 Reputationelle Grimstrategie

Nun soll Grim abgewandelt werden, so dass eine gerade Anzahl von N Spielern zufällig gegen einander antretten. EinE SpielerIn behält ihren/ seinen guten Ruf, wenn sie /er mit Spielern eines guten Rufes kooperiert und gegen Betrüger schummelt. Wenn alle Spieler sich an diese Strategie halten hat man ein perfektes Nashgleichgweicht mit $\delta \geq \frac{1}{2}$.

Beweis. Die Strafe für betrügen ist die gleiche wie bei der klassischen Grimstrategie. □

2.3 Personalisierte Grimstrategie

In dieser Variante wird der Verlauf unter den Spielern nicht sichtbar. Man kann das Spiel als eine Summe unabhängiger Spiele betrachten und erhält ein Nashgleichgewicht für $\delta \geq 1 - \frac{1}{2N}$.

Durchnittliche Gewinn beim Betrügen beträgt:
$(1-\delta)(2 + \delta * 2 + \delta^2 * 2 + ...) = (1-\delta)\frac{2}{1-\delta} = 2$

Offensichtlich maximiert diese Strategie den Durchschnittlichen Gewinn.

Ist das N groß genug ist es in dieser Variante vorteilhaft zu Betrügen, da die Wahrscheinlichkeit auf den gleichen Spieler zu treffen klein ist.

3 Gefahren für Reputationssysteme

Es sollen nun drei Angriffsmethoden vorgestellt werden Reputationssysteme anzugreifen. Dazu gehöhren: Whitewashing, Inkorrektes Feedback und Phantom Feedback.

3.1 Whitewashing

Bei dem Spiel mit reputationellem Grim könnten Betrüger bei schlechtem Ruf einfach ihre Identität wechseln und so das Gleichgeweicht stören.

Weitere verfahren um Whitewashing zu verhindern, beinhalten, dass Neuankömmlinge nur gegen Neuankömmlinge spielen dürfen oder dass es schwer ist eine neue Identität zu erzeugen. Beispielsweise könnten sich Teilnehmer über das PostIdent Verfahren am System identifizieren. Auf jeden Fall ist Whitewashing ein Phänomen, dass eine für alle Teinehmer sozial optimale Strategie (stets zu Kooperieren) verhindert und man so zu Kompromissen gezwungen ist.

3.2 Inkorrektes Feedback

In vielen Situationen gibt es nicht genug Feedback oder die Meinungen sind unehrlich oder verzerrt. Eine Variante ist es, Feedback zu Belohnen. Doch dabei kann es leicht dazu kommen das die allgemeine Meinung als Referenz genommen wird und so Herden oder Informationskaskaden entstehen.

Ein Weg dies zu vermeiden ist es, *peer prediction Modell* zu benutzen. Ziel des Systems ist es die ehrliche Meinung (*Truthful revelation*) eines jeden Spielers zu belohnen.

Jeder Spieler hat eine subjektive Wahrnehmung von der Qualität eines Produktes (im diesem vereinfachten Fall nur gut oder schlecht) und objektive Informationen fehlen oft.

Es wird angenommen, dass Spieler simultan eine Entscheidung treffen und zunächst keine Information über die Entscheidung anderer Spieler haben. Eine Bewertung ist möglich durch eine bedingte Wahrscheinlichkeitsfunktion mit fest gespeicherten Werten(, da Bewertungen einer rationalen Grundlage zu Grunde liegen müssen).

Beobachtung 1. *Eine Antwort $\bar{x}^1$ gilt im Vergleich zu anderen Antworten $\bar{x}^{-1}$ für den Spieler als beste, wenn für jedes m gilt:*

$$\forall \hat{x}^i \in S : E_{S^{-1}}[\tau_i(\bar{x}^i_m, \bar{x}^{-1}) \mid S^i = s_m] \geq E_{S^{-1}}[\tau_i(\hat{x}^i, \bar{x}^{-1}) \mid S^i = s_m]$$

$\bar{x}$ ist ein Nashgleichgweicht im *Simultaneous Reporting* Spiel, wenn die Formel für alle $i = 1..I$ gilt. Sie hat ein striktes Nashgleichgewicht, wenn die obige Ungleichung strikt ist . *Truthful revelation*

ist ein Nashgleichgweicht vom *Simultaneous Reporting* Spiel., wenn die Ungleichung für alle i und m mit $x_m^i = s_m$ erfüllt ist.

Ein Problem ist, dass *Truthful revelation* nicht das einzige Gleichgewicht für das *Simultaneous Reporting* Spiel ist. Zwei weitere wären, wenn entweder alle Spieler immer die Antwort "gutöder immer alle die Antwort ßchlecht"geben. Diese "falschen"Gleichgewichte müssten vom System auf andere Weise erkannt werden.

3.3 Phantom Feedback

Um das Phänomen von Phantom Feedback zu Verstehen müssen wir zunächst annehmen, dass vielen Fällen es kein Objektives Feedback gibt. Der Ruf eines/ einer Spielers/ Spielerin beruht nicht auf dem Feedback sondern seinen Handlungen und/oder Bekannschaften, also der Einschätzung durch andere SpielerInnen. Es entsteht eine Menge von transitiven Vertauensbeziehungen in die SpielerInnen mit einer guten Bewertung durch SpielerInnen aus der vertrauenswürdigen Menge hinzugefügt werden können. Das Problem ist am besten als Graph zu Modellieren: $G = (V, E, t)$:

- V - Menge der Spieler
- E - Menge gerichteter Kannten
- Gewichten: $t : E \to \mathbb{R}^+$
- Bewertungsfunktion: $F : G \to \mathbb{R}^{|V|}$
- $F_v(G)$ Ruf eines Spielers $v \in V$

Phantom Feedback, auch Sybil Attack oder Sock Puppet gennant, tritt auf, wenn ein Spieler falsche Identitäten erstellt um den Ruf seiner/iherer reale Identität zu verbesern. Im Graphen kann dieser Angriff wie folgt definiert werden.

Definition 2. *Sybil Attack tritt auf, wenn gegeben von einem Graph $G = (V, E, t)$ ein Graph $G' = (V', E', t')$ mit Untergraphen $U' \subseteq V'$ erstellt wird. Wenn $v \in U'$ und U' zu v kolabiert, so dass G' mit G gleich wird.*

Definition 3. *Eine Funktion F ist value sybil proof, wenn gilt:*
$\forall G \wedge v \in V$: Es existiert keine Strategie für v, (G',U'), wo für ein beliebiges $u \in U'$ die Ungleichung $F_u(G) > F_v(G)$ erfüllt wäre.

Definition 4. *Eine Funktion F ist rank sybil proof, wenn gilt:*
$\forall G \wedge v \in V$: Es existiert keine Strategie (G',U') für v, wo: $u \in U' \wedge w \in V \backslash \{v\}$: $F_u(G') \geq F_w(G') \wedge F_v(G) < F_w(G)$ zugleich erfüllt sind.

Beobachtung 5. *Symetrische Repuationsfunktionen sind nicht sicher vor rank sybil proof Angriffen.*

Beweis. Diese Beobachtung lässt sich durch Spiegelung des Graphen im Knoten v zeigen. Wenn $F_w(G) > F_v(G)$ gilt, muss es durch Symmetrie einen Knoten u geben, wo $F_u(G') = F_w(G')$ erfüllt ist. Somit sind Symmetrische Funktionen nicht vor *rank sybil proof* Angriffen sicher. □

Mit Min-Cut kann man zeigen, dass Funktionen, die Maximum Flow anwenden sicher vor *value sybil proof* Angriffen sind, da durch hinzufügen von Knoten v seinen Fluss nicht verkleinern kann, aber nicht sicher vor *rank sybil proof* Angriffen. So kann v denn Fluss zu einem Knoten w beinflussen, in dem er eine Sybil der auf dem Maximum Flow Pfad zu w ist wegnimmt.

Dahingegen sind Funktionen vom Typ Pathrank sicher vor allen Sybil Angriffen, da das hinzufügen von Knoten nicht die Distanz zu einem Knoten s_0 verringern kann. Auch kann durch wegnamhe von Sybils nicht die Ungleichung $F_v(G) > F_w(G)$ verändert werden.

4 Zusammenfassung und Ausblick

Es wurde die tägliche (oft unbewusste) Verwendung von Reputationssystemen gezeigt. Mögliche Gefahren wie Whitewashing, inkorektes Feedback und Phantom Feedback erläutert und wie mit Algorithmen diese Gefahren beseitigt werden. So lange man durch Reputationssysteme enkonomische oder soziale Vorteile erhält, wird es stets Angriffe darauf geben und nach Lösungen gesucht diese zu bannen. Als Beispiel aus der Geschichte sei hier Google PageRank erwähnt und zahlreiche verfahren diesen zu Umgehen (sei es durch Linkfarmen = Syil attack) oder Einkaufen von Links.

Bei diesen Systemen gibt es eine vielzahl offener Probleme:

- Umgebungen wo eine neutrale Authorität zum berechnen von Bewertungen fehlt.
- Spieler können in der Lage sein Nachrichten abzufangen und zu verändern.
- Reputationsysteme sollten daher auf Verteilten Systemen implementiert werden.
- Der zitierte Pagerank hat seit 2007 an Bedeutung verloren, neuere Verfahren (seit 2011) beinhalten Trust Rank von Yahoo und Google. Es wäre sinnvoll diesen auf algorithmische Schwächen und Angriffspunkte zu untersuchen.
- Die Autoren der Strategien [1] bemängeln, dass es in der Praxis wenig Systeme gibt, die alle o.g. Sicherheitsmechanismen implementieren.

5 Quellen

[1] Nisan, Roughgarden, Tardos, Vazirani. Algorithmic Game Theory, Kapitel 27