Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **415,** 1445–1468 (2011) doi:10.1111/j.1365-2966.2011.18797.x

Edge modes in self-gravitating disc–planet interactions

Min-Kai Lin★ and John C. B. Papaloizou★

Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA

Accepted 2011 March 25. Received 2011 March 24; in original form 2010 December 6

ABSTRACT

We study the stability of gaps opened by a giant planet in a self-gravitating protoplanetary disc. We find a linear instability associated with both the self-gravity of the disc and local vortensity maxima which coincide with gap edges. For our models, these edge modes develop and extend to twice the orbital radius of a Saturn mass planet in discs with total masses $M_{\rm d} \gtrsim 0.06M_*$, where M_* is the central stellar mass, corresponding to a Toomre $Q \lesssim 1.5$ at twice the planet's orbital radius. The disc models, although massive, are such that they are stable in the absence of the planet. Unlike the previously studied local vortex forming instabilities associated with gap edges in weakly or non-self-gravitating discs with low viscosity, the edge modes we consider are global and exist only in sufficiently massive discs, but for the typical viscosity values adopted for protoplanetary discs.

It is shown through analytic modelling and linear calculations that edge modes may be interpreted as a localized disturbance associated with a gap edge inducing activity in the extended disc, through the launching of density waves excited through gravitational potential perturbation at Lindblad resonances. We also perform hydrodynamic simulations in order to investigate the evolution of edge modes in the linear and non-linear regimes in disc–planet systems. The form and growth rates of developing unstable modes are found to be consistent with linear theory. Their dependence on viscosity and gravitational softening is also explored.

We also performed a first study of the effect of edge modes on disc–planet torques and the orbital migration of the planet. We found that if edge modes develop, then the average torque on the planet becomes more positive with increasing disc mass. In simulations where the planet was allowed to migrate, although a fast type III migration could be seen that was similar to that seen in non-self-gravitating discs, we found that it was possible for the planet to interact gravitationally with the spiral arms associated with an edge mode and that this could result in the planet being scattered outwards. Thus orbital migration is likely to be complex and non-monotonic in massive discs of the type we consider.

Key words: planets and satellites: formation – planet–disc interactions – protoplanetary discs.

1 INTRODUCTION

Understanding the interaction between gaseous protoplanetary discs and embedded planets is important for planet formation theory. Such interaction may lead to inward orbital migration and account for at least some of the observed class of exoplanets called 'hot Jupiters' (Mayor & Queloz 1995) which orbit close to their central stars. Analytical studies of disc–planet interaction began well before the discovery of extrasolar planets (Goldreich & Tremaine 1979). It is known that giant planets with masses comparable to or exceeding that of Saturn can open gaps in standard model discs (Papaloizou & Lin 1984) subsequent to which they may migrate inwards.

The stability of protoplanetary discs with gaps opened by interaction with a planet has also been the subject of study (Koller, Li & Lin 2003; Li et al. 2005; de Val-Borro et al. 2007), as well as the consequences of instability on planetary migration (Ou et al. 2007; Li et al. 2009; Lin & Papaloizou 2010; Yu et al. 2010). These works focused on low-viscosity discs where gap edges become unstable with the result that vortices form. Such instabilities are known to be associated with steep surface density gradients or narrow rings (Papaloizou & Pringle 1985; Lovelace et al. 1999; Li et al. 2000, 2001). These works, like most disc–planet studies, either ignore self-gravity completely or at best consider weak self-gravity. We remark that partial disc gaps induced by a Saturn mass planet may

★E-mail: mkl23@cam.ac.uk (MKL); J.C.B.Papaloizou@damtp.cam.ac.uk (JCBP)

be associated with rapid type III migration (Masset & Papaloizou 2003; Pepliński, Artymowicz & Mellema 2008) in a massive enough disc which is viscous enough to be stable against vortex formation. Lin & Papaloizou (2010) found that type III migration, although unsteady, could on average persist when the disc viscosity was small enough for vortex instability to occur.

However, as type III migration occurs in massive discs the effects of self-gravity need to be properly considered for consistency. The stability of a disc gap induced by interaction with a giant planet in massive, self-gravitating (SG) discs has not yet been studied. The stability of structured SG discs without planets was explored for particle discs by Sellwood & Kahn (1991) and for gaseous discs by Papaloizou & Lin (1989) and Papaloizou & Savonije (1991). Similarly, more recently Meschiari & Laughlin (2008) also demonstrated gravitational instabilities associated with prescribed surface density profiles that model gap structure.

In this paper, we extend the above works by studying the gravitational stability of gaps self-consistently opened by a planet. In this respect the disc models are stable if the planet is not introduced and there is thus no gap. We extend the analytic discussion of neutral modes associated with surface density gap edges given by Sellwood & Kahn (1991) to gaseous discs. In addition we develop the physical interpretation of the associated gap edge instabilities in massive discs as disturbances localized around a vortensity maximum that further perturb gravitationally the smooth parts of the disc by exciting waves at Lindblad resonances. The angular momentum carried away reacts back on the edge disturbance so as to destabilize it.

We perform both linear calculations and non-linear simulations for discs with a range of masses that consistently identify the growth of low azimuthal mode number, m, edge modes as being the dominant form of instability. In addition we evaluate the effect of these on the disc torques acting on the planet. We find that these torques become unsteady and oscillate in time. First estimates of the effect on the planetary migration show that although fast type III like inward migration may occur in the unstable massive discs we study, scattering by spiral arms may also occur leading to short periods of significant outward migration.

This paper is organized as follows. We present the governing equations and basic model in Section 2. In Section 3 we then present the results of an illustrative simulation of an SG disc with an embedded Saturn mass planet that produces a dip/gap in the surface density profile, and subsequently undergoes an instability in which the gap edges play an important role. In the absence of the planet and gap, the Toomre Q value is high enough for the disc model to remain stable.

In Section 4 we present an analytic discussion of the linearly unstable modes. We derive the general wave action conservation law for these modes and apply it to study their angular momentum balance. In particular we show that, when the equation of state is barotropic, at marginal stability the angular momentum loss through waves propagating out of the system is balanced by corotation torques exerted on the disc. These are expected to be strongest near an edge. In Section 5 we consider modes localized around a vortensity maximum near an edge for which the self-gravity response balances the potentially singular response at corotation, showing that such disturbances may be driven by wave excitation at Lindblad resonances. In Section 6 we go on to present linear calculations for various disc models which confirm the existence of edge-dominated modes for low values of the azimuthal mode number. In addition we find instabilities for larger values of m for which edge effects are less dominant, but these modes have weaker growth and do not appear in non-linear numerical simulations.

In Section 7 we present results from hydrodynamic simulations for a range of disc masses. These are all stable in the absence of the planet. However, they exhibit low m edge-dominated modes once a planetary induced gap is present. The form and behaviour of these are found to be in accord with linear theory. However, in the simulations with the lower Q values, the gap is found to widen and deepen as the simulation progresses. In addition, the global angular momentum transport through the disc measured through an effective α parameter is increased above the level that would be induced by the planet alone. In addition the spiral arms associated with the edge instabilities are shown to produce fluctuating torques acting on the planet. Fast inwards type III migration may occur, but we also observed *outward* migration due to interaction with spiral arms. This indicates that migration is unlikely to be a simple monotonic process in a gravitationally active disc. Finally in Section 9 we summarize our results and conclude.

2 BASIC EQUATIONS AND MODEL

We describe here the governing equations and models for the disc–planet system as used in analytic discussions, linear calculations and solved in hydrodynamic simulations. The system we consider is a gaseous disc of mass $M_{\rm d}$ orbiting a central star of mass M_*. We adopt a cylindrical, non-rotating, coordinate system (r, φ, z) centred on the star, where z is the vertical coordinate increasing in the direction normal to the disc mid-plane. It is convenient to adopt a system of equations in three dimensions to begin the discussion of angular momentum conservation in Section 4.1 below. We begin by listing these and then go on to explain how we adapt them to obtain the system governing a two-dimensional razor thin disc that we solve in hydrodynamic simulations. They are the continuity equation :

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = 0, \tag{1}$$

and the equation of motion

$$\rho \left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} \right) = -\nabla p - \rho \nabla \Phi - \rho \nabla \Phi_{\rm ext} + \boldsymbol{f}. \tag{2}$$

Here ρ is the density, $\boldsymbol{u}$ is the velocity field, Φ is the gravitational potential due to the disc, $\Phi_{\rm ext}$ is the potential due to external bodies, p is the pressure and $\boldsymbol{f}$ is the viscous force. The gravitational potential due to the disc satisfies Poisson's equation:

$$\nabla^2 \Phi = 4\pi G \rho, \tag{3}$$

which yields the integral expression

$$\Phi = -G \int_{\mathcal{D}} \frac{\rho(r', \varphi', z') r' {\rm d}r' {\rm d}\varphi' {\rm d}z'}{\sqrt{r^2 + r'^2 - 2rr' \cos(\varphi - \varphi') + (z - z')^2}}, \tag{4}$$

where $\mathcal{D}$ is the domain where ρ is non-zero. This coincides with the disc domain when this is not separated from external material.

For analytic discussion we consider fluids with a general barotropic equation of state for which $p = p(\rho)$ and ${\rm d}p/{\rm d}\rho = c_{\rm s}^2$, with $c_{\rm s}$ being the local sound speed; for numerical calculations a locally isothermal equation of state is adopted, $p = c_{\rm s}^2 \rho$, where $c_{\rm s}(r, z)$ is a specified function of r and z.

2.1 Razor thin limit

To obtain the limiting case of a razor thin disc, we assume no interior vertical motions and that the radial and azimuthal velocity components are independent of z. We then integrate over z so that ρ

is replaced by the surface density Σ in equations (1) and (2) while p becomes a vertically integrated pressure, which in the barotropic case is assumed to be a function only of Σ with $\mathrm{d}p/\mathrm{d}\Sigma = c_\mathrm{s}^2$. The z dependence of the potentials is neglected so that we now have the governing equations

$$\frac{\partial \Sigma}{\partial t} + \nabla \cdot (\boldsymbol{u}\Sigma) = 0, \tag{5}$$

and

$$\Sigma\left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right) = -\nabla p - \Sigma\nabla\Phi - \Sigma\nabla\Phi_\mathrm{ext} + \boldsymbol{f}, \tag{6}$$

where $\boldsymbol{u} \equiv (u_\mathrm{r}, u_\varphi)$ is now a two-dimensional velocity field and $\boldsymbol{f}$ is now the two-dimensional viscous force, characterized by a uniform kinematic viscosity ν in our models (see Masset 2002, for details). The mid-plane disc potential is given by

$$\Phi = -\int_\mathcal{D} \frac{G\Sigma(r',\varphi')}{\sqrt{r^2 + r'^2 - 2rr'\cos(\varphi-\varphi') + \epsilon_\mathrm{g}^2}} r'\mathrm{d}r'\mathrm{d}\varphi', \tag{7}$$

where we have introduced a softening length $\epsilon_\mathrm{g} = \epsilon_{\mathrm{g}0}H(r')$, with $\epsilon_{\mathrm{g}0}$ being a dimensionless constant and a putative semithickness for razor thin discs defined through $H(r) \equiv hr = c_\mathrm{s}/\Omega_\mathrm{k}$, where $\Omega_\mathrm{k} = \sqrt{GM_*/r^3}$ is the Keplerian rotation rate. The dimensionless constant disc aspect ratio is h.

The gravitational potential due to external bodies comprising the central star and an embedded planet is

$$\begin{aligned}\Phi_\mathrm{ext} = &-\frac{GM_*}{r} - \frac{GM_\mathrm{p}}{\sqrt{r^2 + r_\mathrm{p}^2 - 2rr_\mathrm{p}\cos(\varphi-\varphi_\mathrm{p}) + \epsilon_\mathrm{p}^2}} \\ &+ r\int_\mathcal{D} \frac{G\Sigma(r',\varphi')}{r'^2}\cos(\varphi-\varphi')r'\mathrm{d}r'\mathrm{d}\varphi' \\ &+ \frac{GM_\mathrm{p}}{r_\mathrm{p}^2} r\cos(\varphi-\varphi_\mathrm{p}).\end{aligned} \tag{8}$$

Here M_p is the fixed mass of the embedded planet with cylindrical coordinates $[r_\mathrm{p}(t), \varphi_\mathrm{p}(t)]$. The associated gravitational potential is softened with softening length $\epsilon_\mathrm{p} = \epsilon_{\mathrm{p}0}H(r_\mathrm{p})$, where $\epsilon_{\mathrm{p}0}$ is a dimensionless constant. The last two terms in the expression for Φ_ext which give the indirect potential account for the forces due to the disc and embedded planet acting on the central star. They occur because we adopt a non-inertial frame of reference.

2.2 Model setup

We describe specific disc and planet models used in numerical calculations, which also motivate the analytical discussion below. We adopt units such that $G = M_* = 1$ and the inner boundary radius of the disc, r_i, is unity. The Keplerian orbital period at this radius $r = r_\mathrm{i} = 1$ is then $P(1) = 2\pi$. The disc occupies $r = [r_\mathrm{i}, r_\mathrm{o}] = [1, 10]$. The disc is taken to have a locally isothermal equation of state, $p = c_\mathrm{s}^2\Sigma$, where $c_\mathrm{s} = rh\Omega_\mathrm{k}$. We remark that softening is introduced to take account of the vertical thickness of the disc. Typically, softening length to semithickness ratios of 0.3–0.6 are used (e.g. Masset 2002; Baruteau & Masset 2008). Thus fiducial values for the constants $\epsilon_{\mathrm{p}0}$, $\epsilon_{\mathrm{g}0}$ and h of 0.6, 0.3 and 0.05 are respectively adopted. The initial surface density profile is taken to be given by

$$\Sigma(r) = \Sigma_0 r^{-3/2}\left(1 - \sqrt{\frac{r_\mathrm{i}}{r+H_\mathrm{i}}}\right), \tag{9}$$

(see Armitage & Hansen 1999) where $H_\mathrm{i} = H(r_\mathrm{i})$. This form for Σ, which vanishes smoothly a distance H_i inside the disc inner boundary, has been introduced to ensure that both the gravitational force and the pressure gradient in hydrostatic equilibrium at the disc inner boundary give minor contributions and vary smoothly there. The constant surface density scale Σ_0 is adjusted so that $Q(r_\mathrm{o})$, where

$$Q(r) = \frac{c_\mathrm{s}\kappa}{\pi G\Sigma} \rightarrow \frac{hM_*}{\pi r^2\Sigma(r)} \tag{10}$$

is the Toomre Q parameter, evaluated in the limit of a thin Keplerian disc for which the epicycle frequency $\kappa = \Omega_\mathrm{k}$, takes on a specified value Q_o. Thus Q_o parametrizes the models.

The initial azimuthal velocity is found by assuming hydrostatic equilibrium in which the centrifugal force balances forces due to stellar gravity and the disc's self-gravity and pressure gradient. Thus the initial azimuthal velocity is given by

$$u_\varphi^2 = \frac{r}{\Sigma}\frac{\mathrm{d}p}{\mathrm{d}r} + \frac{GM_*}{r} + r\frac{\mathrm{d}\Phi}{\mathrm{d}r}, \tag{11}$$

For the local isothermal equation of state we have adopted, the contribution due to the pressure is given by

$$\frac{r}{\Sigma}\frac{\mathrm{d}p}{\mathrm{d}r} = c_\mathrm{s}^2\left\{-\frac{5}{2} + \frac{r\sqrt{r_\mathrm{i}}(r+H_\mathrm{i})^{-3/2}}{2\left[1-\sqrt{r_\mathrm{i}/(r+H_\mathrm{i})}\right]}\right\}. \tag{12}$$

At $r = r_\mathrm{i}$, for $h = 0.05$, this is $\sim 20c_\mathrm{s}^2$ which is approximately 5 per cent of the square of the local Keplerian speed so that the contribution of the pressure force is indeed minor when compared to that arising from the gravity of the central star.

The initial radial velocity is set to $u_\mathrm{r} = 3\nu/r$, corresponding to the initial radial velocity of a one-dimensional Keplerian accretion disc with $\Sigma \propto r^{-3/2}$ and uniform kinematic viscosity. The disc is evolved for a time $\sim 280P(r_\mathrm{i})$ before introducing the planet.

When a planet of mass M_p is introduced, it is inserted in a circular orbit, under the gravity of both the central star and disc, at a distance $r_\mathrm{p} = r_\mathrm{p}(t=0)$ from the central star. We quote time in units of the Keplerian orbital period at the planet's initial radius, which is given by $P_0 = 2\pi/\Omega_\mathrm{k}[r_\mathrm{p}(t=0)]$. If the planet is held on fixed circular orbit, then $r_\mathrm{p}(t) = r_\mathrm{p}(t=0)$.

3 NUMERICAL SIMULATIONS

In order to motivate and facilitate our analytic discussion, linear analysis and interpretation of the instabilities we find in an SG disc with a surface density gap or dip induced by a massive planet, we here provide a brief demonstration of their existence. We show global instabilities associated with a surface density depression made self-consistently by a giant planet in a fixed circular orbit by means of numerical simulations. Details of the numerical approach are given in Section 7 where additional hydrodynamic simulations of this type as well as others, where the planet is allowed to migrate, will be investigated more fully.

3.1 The nature of edge modes at the edges of disc surface density depressions induced by interaction with giant planets

Here, we describe results obtained for a disc with the initial density profile specified above in which there is an embedded planet with mass $M_\mathrm{p} = 3\times10^{-4}M_*$. This corresponds to a Saturn mass planet when $M_* = 1\,\mathrm{M}_\odot$. The disc model is $Q_\mathrm{o} = 1.5$, corresponding to total disc mass of $M_\mathrm{d} = 0.063M_*$. The uniform kinematic viscosity was taken to be $\nu = 10^{-5}$ in code units. The planet was introduced at radius $r_\mathrm{p} = 5$ and at $t = 25P(r_\mathrm{p}) \equiv 25P_0$. Its gravitational potential was then ramped up over a time interval $10P_0$. For this simulation it

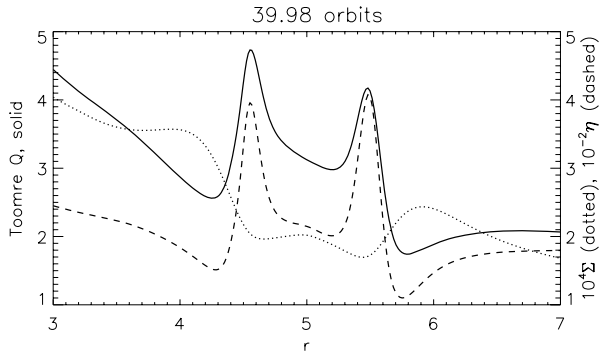

Figure 1. Gap profile produced by a Saturn mass planet embedded in a disc with $Q_o = 1.5$. The azimuthally averaged surface density (dotted line), vortensity (dashed line) and Toomre Q parameter (solid line) are plotted. Note that the Q profile shows a maximum and minimum associated with each gap edge. The planet is in a fixed circular orbit located at $r = 5$.

was held on a fixed circular orbit. Without a perturber we expect, and find, the disc to be gravitationally stable because we have $Q(r_p) = 2.62$ and near the outer boundary $Q \sim 1.5$.

The profile of the gap opened by the planet is shown in Fig. 1. The azimuthally averaged surface density $\langle\Sigma\rangle_\varphi$, vortensity $\langle\eta\rangle_\varphi$ where $\eta \equiv \kappa^2/2\Omega\Sigma$, and Toomre Q value are plotted. The latter is calculated using the azimuthally averaged epicycle frequency. We remark that for an axisymmetric disc, the Toomre parameter is proportional to the product of the ratio of the rotation frequency Ω to the epicycle frequency κ and the vortensity η. It is seen that there is a surface density depression, referred to a gap, associated with a decrease of surface density by $\simeq$ 20 per cent relative to the unperturbed disc.

It has been found that extrema in the vortensity are associated with instability (Papaloizou & Lin 1989). For typical disc models structured by disc–planet interactions, such as those illustrated in Fig. 1, local maxima/minima in Q and η have been found to approximately coincide. The neighbourhoods of the inner and outer gap edges both contain maxima and minima of Q and η.

In the absence of a planet, Q smoothly decreases outwards. In the case when a planet is present, disc–planet interaction results in significant vortensity generation as material flows through shocks (Lin & Papaloizou 2010), leading to vortensity maxima. The resulting Q and $\langle\eta\rangle_\varphi$ profiles are very similar [since $Q = (c_s/\pi G)(2\Omega\eta/\Sigma)^{1/2}$]. Fig. 1 shows that the planet-modified Q profile may exhibit a range of behaviours close to r_p. Vortensity diffusion, e.g. due to significantly larger viscosities than those considered here, could render the Q profile to be uniform. In such cases, unstable modes associated with vortensity maxima cannot be set up. The unperturbed Toomre Q profile (prior to the planet introduction) may also play a role. The Toomre Q profile perturbed by the planet is reminiscent of its unperturbed value, without planet. If the latter happens to be approximately uniform, the net impact of edge modes on the planet torque could be negligible. In our models, the background Toomre Q decreases with radius, so we expect that the outer gap edge is more gravitationally unstable than the inner gap edge.

Returning to the present case, Fig. 1 shows that in the neighbourhood of the outer gap edge, the maximum and minimum values of Q occur at $r = 5.45$ and 5.75 are $Q = 4.2$ and 1.75, respectively. In the region of the inner gap edge, the maximum and minimum values of Q occur at $r = 4.55$ and 4.25 are $Q = 4.75$ and 2.55, respectively. The extrema are separated by $\simeq 1.4H$ and $1.1H$ in the inner and outer gap edge regions, respectively. Thus the characteristic width of the gap edge is the local scaleheight. On average $Q \sim 2$ for $r > 6$. Since $Q > 1$ everywhere, the disc is stable against local axisymmetric perturbations.

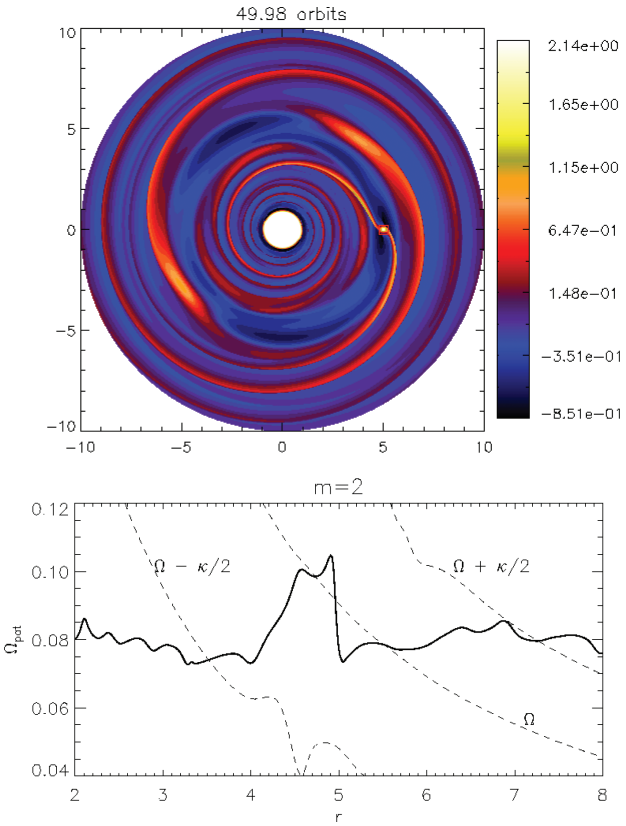

Figure 2. Relative surface density perturbation for $Q_o = 1.5$ at $t = 50P_0$ (top) and the radial form of the pattern speed found from the $m = 2$ component of the Fourier transform of the surface density (bottom, solid line). Note that this mode is associated with the outer disc and, apart from within the gap, this mode appears to have a stable pattern speed of $\Omega_{pat} \sim 0.078$ corresponding to a corotation point at $r = 5.5$. Plots of Ω and $\Omega \pm \kappa/2$ are also given (bottom, dashed line).

The simulation shows that both gap edges become unstable. The surface density contours at $t = 50P_0$ are shown in Fig. 2. We identify a mode with $m = 3$ and relative density perturbation $\Delta\Sigma/\Sigma \simeq 0.4$ associated with the inner edge and a mode with $m = 2$ and $\Delta\Sigma/\Sigma \simeq 0.9$ associated with the outer edge. The modes have become non-linear, with the development of shocks, on dynamical timescales. Spirals do not extend across the gap, indicating effective gap opening by the planet and disc self-gravity is not strong enough to connect the spiral modes on either side of r_p.

Plots of the radial dependence of the pattern rotation speed, Ω_{pat}, obtained from the $m = 2$ component of the Fourier transform of the surface density, together with the azimuthally averaged values of Ω and $\Omega \pm \kappa/2$ are presented in Fig. 2. We focus on the mode associated with the outer gap edge because the $m = 2$ spiral mode has more than twice the relative surface density amplitude compared to the inner spiral mode.

By necessity, the Fourier transform includes features due to the planetary wakes and inner disc spiral modes as well as the dominant outer disc mode. On average, $\Omega_{pat} \sim 0.08$.[1] There is a corotation

[1] We have explicitly checked that this is the pattern speed by measuring the angle through which the spiral pattern rotates in a given time interval.

point in the outer disc at $r = 5.5$ where $\Omega = \Omega_{\rm pat} \sim 0.078$, i.e. at the local vortensity (surface density) maximum (minimum) of the basic state, which is just within the gap (see Fig. 1). This is consistent with pattern speeds obtained from linear calculations presented later. Fig. 2 indicates another corotation point at $r \simeq 4.7$, but this is due to inclusion of features associated with the planetary wakes.

We refer to unstable spiral modes identified here as *edge modes*. Their properties can be compared to those of *groove modes* in particle discs (Sellwood & Kahn 1991) or fluid discs (Meschiari & Laughlin 2008). Sellwood & Kahn describe groove modes as gravitational instabilities associated with the coupling of disturbances associated with two edges across the gap between them. However, our analytic description of the edge mode instability is a coupling between a disturbance at a single gap edge and the disturbance it excites in the adjoining smooth disc away from corotation. The other gap edge plays no role. Although both types of mode are associated with vortensity maxima, the groove modes described above would have corotation at the gap centre midway between the edges, whereas in our case corotation is at the gap edge.

This is significant because in our model there will be differential rotation between the spiral pattern and the planet. In addition, unstable modes on the inner and outer gap edges need not have the same m (Fig. 2 shows $m = 3$ on the inner edge and $m = 2$ on the outer edge). If there were a groove mode with corotation at the gap centre, then the spiral pattern would corotate with the planet and the coupled edges must have disturbances dominated by the same value of m.

3.2 Comparison to vortex instabilities

We have demonstrated a global instability in SG disc–planet systems with a gap. It is interesting to compare these spiral modes to the well-known localized vortex-forming instabilities that occur in non-self-gravitating (NSG) or weakly self-gravitating (SG) discs near gap edges (Li et al. 2009; Lin & Papaloizou 2010). However, the vortex instability requires low viscosity. For comparison purposes, here we use a kinematic viscosity of $\nu = 10^{-6}$ [or an α viscosity parameter $O(10^{-4})$]. This value of ν is an order of magnitude smaller than what has been typically adopted for protoplanetary disc models. We compare the behaviour of disc models with $Q_{\rm o} = 1.5$ and 4.0, the former having strong self-gravity and the latter weak self-gravity.

Fig. 3 shows two types of instabilities depending on disc mass. The lower mass disc with $Q_{\rm o} = 4$ develops six vortices localized in the vicinity of the outer gap edge. The more massive disc with $Q_{\rm o} = 1.5$ develops edge modes of the type identified in the earlier run with $\nu = 10^{-5}$. Here, the $m = 3$ spiral mode is favoured because of the smaller viscosity coefficient used.

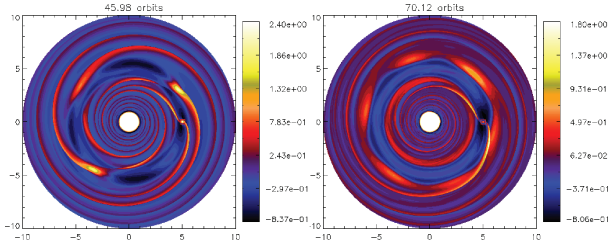

Figure 3. Relative surface density perturbation for the $Q_{\rm o} = 1.5$ disc (left) and the $Q_{\rm o} = 4.0$ disc (right) with $\nu = 10^{-6}$. These plots show that the gap opened by a Saturn mass planet supports vortex instabilities in low-mass discs and spiral instabilities in sufficiently massive discs.

Vortex modes are said to be local because instability is associated with flow in the vicinity of corotation (Lovelace et al. 1999) and does not require the excitation of waves. Edge modes are said to be global because instability requires interaction between an edge disturbance and waves launched at Lindblad resonances, away from corotation, in the smooth part of the disc. Vortices perturb the disc even without self-gravity (Paardekooper, Lesur & Papaloizou 2010). Although the vortex mode in Fig. 3 shows significant wave perturbations in the outer parts of the disc, these should be seen as a consequence of unstable vortices forming around corotation, rather than the cause of a linear instability.

We found the vortex modes have corotation close to local vortensity *minima* in the undisturbed gap profile. The vortex modes are localized with high m being dominant ($m > 5$). On the other hand, the spiral modes found here are global with low $m < 5$. Edge modes make the gap less identifiable than vortex modes do because the former, with corotation closer to $r_{\rm p}$, protrudes the gap edge more. While the edge mode only takes a few P_0 to become non-linear with spiral shocks attaining comparable amplitudes to the planetary wakes, the vortex mode takes a significantly longer time (the plot for $Q_{\rm o} = 4$ is $30P_0$ after gap formation).

4 ANALYTICAL DISCUSSION OF EDGE MODES

The fiducial disc simulated above is massive with $M_{\rm d} \sim 0.06 M_*$ but it is still stable against gravitational instabilities in the sense that $Q \geq 1.5$ everywhere, and the disc without an embedded planet, which has a smoothly varying surface density profile, does not exhibit the spontaneous growth of spiral instabilities. When a planet of Saturn's mass is inserted ($M_{\rm p} = 3 \times 10^{-4} M_*$), it is expected to only open a partial gap (∼30 per cent deficit in surface density), but even this is sufficient to trigger a $m = 2$ spiral instability.

The fiducial case above thus suggests spiral modes may develop under conditions that are not as extreme as those considered by Meschiari & Laughlin (2008). They used a prescribed gap profile with a gap depth of 90 per cent relative to the background, which is three times deeper than in our models. Their gap corresponds to $M_{\rm p} = 0.002 M_*$, although a planet potential was not explicitly included. In both their model and ours, the gap width is $\simeq 2 r_{\rm h}$, where $r_{\rm h} = (M_{\rm p}/3M_*)^{1/3} r_{\rm p}$ is the Hill radius, but since they effectively used a two Jupiter mass planet, at the same $r_{\rm p}$ their gap is about 1.9 times wider than ours.

We devote this section and the next to a theoretical discussion of spiral modes associated with planetary gap edges. This work is based on an analysis of the governing equations for linear perturbations. As angular momentum balance is important for enabling small perturbations to grow unstably, we begin by formulating the conservation of angular momentum for linear perturbations.

4.1 The conservation of angular momentum for a perturbed disc

We derive a conservation law for the angular momentum associated with the perturbations of a disc that enables the angular momentum density and flux to be identified within the framework of linear perturbation theory. The behaviour of these quantities is found to be important for indicating the nature of the angular momentum balance in a system with a neutral or weakly growing normal mode and how positive (negative) angular momentum fluxes associated with wave losses may drive the instability of a disturbance that decreases (increases) the local angular momentum density.

4.2 Barotropic discs

We begin by writing down the linearized equation of motion in three dimensions for the Lagrangian displacement $\boldsymbol{\xi} \equiv (\xi_r, \xi_\varphi, \xi_z)$ for a differentially rotating fluid with a barotropic equation of state and with self-gravity included (see e.g. Lynden-Bell & Ostriker 1967; Lin, Papaloizou & Kley 1993) in the form:

$$\frac{\mathrm{D}^2\boldsymbol{\xi}}{\mathrm{D}t^2} + 2\Omega\hat{\boldsymbol{k}} \wedge \frac{\mathrm{D}\boldsymbol{\xi}}{\mathrm{D}t} + r\hat{\boldsymbol{r}}(\boldsymbol{\xi}\cdot\nabla\Omega^2) = -\nabla S' - \nabla\Phi'_{\rm ext}, \tag{13}$$

where

$$S' = c_{\rm s}^2\frac{\rho'}{\rho} + \Phi'. \tag{14}$$

Here we adopt a cylindrical polar coordinate system (r, φ, z), perturbations to quantities are denoted with a prime while unprimed quantities refer to the background state, the operator $\mathrm{D}/\mathrm{D}t \equiv \partial/\partial t + \Omega\partial/\partial\varphi$ is the convective derivative following the unperturbed motion and $\hat{\boldsymbol{k}}$ is the unit vector in the z direction which is normal to the disc mid-plane. For perturbations that depend on φ through a factor $\exp(\mathrm{i}m\varphi)$, m being the azimuthal mode number, assumed positive, that we consider, the operator $\mathrm{D}/\mathrm{D}t$ reduces to the operator $(\partial/\partial t + \mathrm{i}m\Omega)$. In addition we recall that for a barotropic equation of state, $\Omega = \Omega(r)$ depends only on the radial coordinate. Then (13) becomes

$$\frac{\partial^2\boldsymbol{\xi}}{\partial t^2} + 2\Omega\hat{\boldsymbol{k}} \wedge \frac{\partial\boldsymbol{\xi}}{\partial t} + 2\mathrm{i}m\Omega\frac{\partial\boldsymbol{\xi}}{\partial t} + 2\mathrm{i}m\Omega^2\hat{\boldsymbol{k}}\wedge\boldsymbol{\xi} + r\hat{\boldsymbol{r}}(\boldsymbol{\xi}\cdot\nabla\Omega^2) - m^2\Omega^2\boldsymbol{\xi} = -\nabla S' - \nabla\Phi'_{\rm ext}. \tag{15}$$

The perturbation to the density is given by

$$\rho' = -\nabla\cdot(\rho\boldsymbol{\xi}). \tag{16}$$

The perturbation to the disc's gravitational potential is given by linearizing Poisson's equation to give

$$\nabla^2\Phi' = 4\pi G\rho'. \tag{17}$$

We have also added an external potential perturbation $\Phi'_{\rm ext}$ to assist with the identification of angular momentum flux later on.

By taking the scalar product of (15) with $\boldsymbol{\xi}^*$, multiplying by the background density, ρ, and taking the imaginary part, after making use of equations (16) and (17) we obtain a conservation law that expresses the conservation of angular momentum for the perturbations in the form:

$$\frac{\partial\rho_J}{\partial t} + \nabla\cdot(\boldsymbol{F}_{\rm A} + \boldsymbol{F}_{\rm G} + \boldsymbol{F}_{\rm ext}) = T, \tag{18}$$

where

$$\rho_J \equiv -\frac{m\rho}{2}\mathrm{Im}\left(\boldsymbol{\xi}^*\cdot\frac{\partial\boldsymbol{\xi}}{\partial t} + \Omega\hat{\boldsymbol{k}}\cdot\boldsymbol{\xi}\wedge\boldsymbol{\xi}^* + \mathrm{i}m\Omega|\boldsymbol{\xi}|^2\right) \tag{19}$$

$$\boldsymbol{F}_{\rm A} = -\frac{m\rho}{2}\mathrm{Im}(\boldsymbol{\xi}^* S') \tag{20}$$

$$\boldsymbol{F}_{\rm G} = -\frac{m}{2}\mathrm{Im}\left(\frac{1}{4\pi G}\Phi'\nabla\Phi'^*\right) \tag{21}$$

$$\boldsymbol{F}_{\rm ext} = -\frac{m\rho}{2}\mathrm{Im}(\boldsymbol{\xi}^*\Phi'_{\rm ext}) \tag{22}$$

and

$$T = \frac{m}{2}\mathrm{Im}(\Phi'_{\rm ext}\rho'^*). \tag{23}$$

Here ρ_J is the angular momentum density, and the angular momentum flux is split into three contributions: $\boldsymbol{F}_{\rm A}$, being proportional to the Lagrangian displacement, is the advective angular momentum flux; $\boldsymbol{F}_{\rm G}$ is the flux associated with the perturbed gravitational stresses and $\boldsymbol{F}_{\rm ext}$ is the flux associated with the external potential which is inactive for free perturbations.

The quantity T is a torque density associated with the external potential as can be seen from that fact that the real torque integrated over azimuth and divided by 2π is

$$-\frac{1}{2\pi}\int_0^{2\pi}\mathrm{Re}(\rho'^*)\mathrm{Re}\left(\frac{\partial\Phi'_{\rm ext}}{\partial\varphi}\right)\mathrm{d}\varphi = T = \frac{m}{2}\mathrm{Im}(\Phi'_{\rm ext}\rho'^*). \tag{24}$$

This justifies the scalings used for the angular momentum density and fluxes.

4.3 The conservation of angular momentum in the two-dimensional razor thin disc limit

The form of the conservation law for a razor thin disc is obtained by integrating equation (18) over the vertical coordinate. For terms $\propto\rho$, the integrand is non-zero only within the disc. $\boldsymbol{\xi}$ and the gravitational potentials may be assumed to depend only on r and φ. The effect of the integration is simply to replace ρ and ρ' by the surface density Σ and its perturbation Σ', respectively. The fluxes become vertically integrated fluxes. For the term associated with the gravitational stresses, assuming the potential perturbation vanishes at large distances, we have

$$\int_{-\infty}^{\infty}\nabla\cdot\boldsymbol{F}_{\rm G}\mathrm{d}z = -\frac{m}{2}\mathrm{Im}[\Phi'(r,\varphi,0)\Sigma'^*] = \frac{1}{r}\frac{\partial(rF_{\rm Gr})}{\partial r} \tag{25}$$

where

$$F_{\rm Gr} = -\frac{m}{8\pi G}\mathrm{Im}\int_{-\infty}^{\infty}\Phi'\frac{\partial\Phi'^*}{\partial r}\,\mathrm{d}z. \tag{26}$$

Thus for a two-dimensional razor thin disc the conservation law is of the same form as (18) but with ρ_J, $\boldsymbol{F}_{\rm A}$ and $\boldsymbol{F}_{\rm ext}$ given by (19), (20) and (22) with ρ replaced by Σ, respectively. As the vertical direction has been integrated over only the radial components of the fluxes contribute. From the above analysis, the vertically integrated angular momentum flux due to gravitational stresses becomes

$$\boldsymbol{F}_{\rm G} \to F_{\rm Gr}\hat{\boldsymbol{r}}, \tag{27}$$

with $F_{\rm Gr}$ given by (26). Accordingly, this term, unlike the others, involves an integration over the vertical direction. We shall obtain explicit expressions for the fluxes in terms of the perturbation $S' = \Sigma' c_{\rm s}^2/\Sigma + \Phi'$ appropriate to the razor thin disc below.

4.4 The conservation of angular momentum for a disc with a locally isothermal equation of state

It is possible to repeat the analysis of Section 4.1 when the disc has a locally isothermal equation of state. In this case, $P = \rho c_{\rm s}^2$, where $c_{\rm s}$ is a prescribed function of position, and the linearized equation of motion (13) is modified to read

$$\frac{\mathrm{D}^2\boldsymbol{\xi}}{\mathrm{D}t^2} + 2\Omega\hat{\boldsymbol{k}}\wedge\frac{\mathrm{D}\boldsymbol{\xi}}{\mathrm{D}t} + r\hat{\boldsymbol{r}}(\boldsymbol{\xi}\cdot\nabla\Omega^2) = -\nabla S' + \frac{\rho'}{\rho}\nabla\left(c_{\rm s}^2\right) - \nabla\Phi'_{\rm ext}. \tag{28}$$

An expression of the form (18) may be derived but now the torque density T takes the form

$$T = \frac{m}{2}\left[\mathrm{Im}(\Phi'_{\rm ext}\rho'^*) - \mathrm{Im}\left(\rho'\boldsymbol{\xi}^*\cdot\nabla\left(c_{\rm s}^2\right)\right)\right]. \tag{29}$$

When $c_{\rm s}$ is constant corresponding to a strictly isothermal equation of state, the fluid is again barotropic and previous expressions

recovered. Equation (29) contains terms in addition to external contributions. In general, these imply the possibility of an exchange of angular momentum between perturbations and the background. A very similar discussion applies to the razor thin limit. We now go on to apply the above analysis in determining the angular momentum balance for the linear perturbations of razor thin discs in more detail and thus determine properties of and conditions for unstable normal modes.

4.5 Properties of the linear modes of a razor thin disc

We assume linear perturbations for which the φ and t dependence is through a factor of the form $e^{\mathrm{i}(\sigma t+m\varphi)}$, where σ is the complex eigenfrequency and m is the azimuthal mode number. From now on this factor is taken as read. Linearizing the hydrodynamic equations (5) and (6) for the inviscid case and ignoring external potentials, we obtain

$$\mathrm{i}\bar{\sigma}\,\Sigma' = -\frac{1}{r}\frac{\mathrm{d}(\Sigma r u'_{\mathrm{r}})}{\mathrm{d}r} - \frac{\mathrm{i}m\Sigma u'_{\varphi}}{r} \tag{30}$$

$$\mathrm{i}\bar{\sigma}u'_{\mathrm{r}} - 2\Omega u'_{\varphi} = -\frac{\mathrm{d}S'}{\mathrm{d}r} + \frac{\Sigma'}{\Sigma}\frac{\mathrm{d}c_{\mathrm{s}}^2}{\mathrm{d}r} \tag{31}$$

$$\mathrm{i}\bar{\sigma}u'_{\varphi} + \frac{\kappa^2}{2\Omega}u'_{\mathrm{r}} = -\frac{\mathrm{i}mS'}{r}, \tag{32}$$

where $\bar{\sigma} = \sigma + m\Omega$ is the shifted mode frequency. Using equations (31) and (32), to eliminate u'_{r} and u'_{φ} in equation (30) we obtain (see e.g. Papaloizou & Savonije 1991)

$$\begin{aligned} r\Sigma' = {} & \frac{\mathrm{d}}{\mathrm{d}r}\left[\frac{r\Sigma}{D}\left(\frac{\mathrm{d}S'}{\mathrm{d}r} + \frac{2m\Omega\bar{\sigma}S'}{r\kappa^2}\right)\right] \\ & - \frac{2m\Omega\bar{\sigma}\Sigma}{\kappa^2 D}\left(\frac{\mathrm{d}S'}{\mathrm{d}r} + \frac{2m\Omega\bar{\sigma}S'}{r\kappa^2}\right) + \frac{mS'}{\bar{\sigma}}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{\eta}\right) \\ & - \frac{4m^2\Omega^2\Sigma S'}{r\kappa^4} - \frac{\mathrm{d}}{\mathrm{d}r}\left[\frac{r\Sigma'}{D}\frac{\mathrm{d}c_{\mathrm{s}}^2}{\mathrm{d}r}\right] + \frac{2m\Omega\Sigma'}{Dr\bar{\sigma}}\frac{\mathrm{d}c_{\mathrm{s}}^2}{\mathrm{d}r}, \end{aligned} \tag{33}$$

where $D \equiv \kappa^2 - \bar{\sigma}^2$, and recall $\eta = \kappa^2/2\Omega\Sigma$ is the vortensity. This form shows that a corotation singularity where $\bar{\sigma} = 0$ can be avoided if corotation corresponds to a vortensity extremum.

The above expressions may be applied to a disc, either with a locally isothermal equation of state, or with a barotropic equation of state, where the integrated pressure $P = P(\Sigma)$ and the sound speed $c_{\mathrm{s}}^2 \equiv \mathrm{d}P/\mathrm{d}\Sigma$. To obtain the latter case, the quantity $\mathrm{d}c_{\mathrm{s}}^2/\mathrm{d}r$ is simply replaced by zero. We remark that inclusion of additional terms dependent on this has been found numerically to produce only a slight modification of the discussion that applies when they are neglected. This is because c_{s}^2 varies slowly compared to either the linear perturbations or the surface density in the vicinity of the edge, thus we shall neglect $(r/c_{\mathrm{s}}^2)\mathrm{d}c_{\mathrm{s}}^2/\mathrm{d}r$ from now on. We also recall that $S' = \Sigma' c_{\mathrm{s}}^2/\Sigma + \Phi'$ and the gravitational potential is given by the vertically integrated Poisson integral as

$$\Phi' = S' - \frac{\Sigma' c_{\mathrm{s}}^2}{\Sigma} = -G\int_{\mathcal{D}} K_m(r, r')\Sigma'(r')r'\mathrm{d}r', \tag{34}$$

where

$$K_m(r, r') = \int_0^{2\pi} \frac{\cos(m\varphi)\mathrm{d}\varphi}{\sqrt{r^2 + r'^2 - 2rr'\cos(\varphi) + \epsilon_{\mathrm{g}}^2}}. \tag{35}$$

Here the domain of integration $\mathcal{D}$ is that of the disc, provided there is no external material. However, in a situation where density waves propagate beyond the disc boundaries, either $\mathcal{D}$ should be formally extended or the form of K_m changed to properly reflect the boundary conditions. We remark that (34) enables Σ' to be determined in terms of S' and if this is used in (33) a single eigenvalue equation for S' results (e.g. Papaloizou & Savonije 1991).

We shall now consider the angular momentum flux balance for normal modes. For simplicity we shall consider the barotropic case. The analytic discussion concerns a disc model with one boundary being a sharp edge with arbitrary propagation of density waves directed away from this boundary being allowed. When applied to a gap opened by a planet in a global disc, the analytic model corresponds the section of the disc from the inner disc boundary to the inner gap edge or the section from the outer gap edge to the outer disc boundary. These two sections are assumed to be decoupled in the analytic discussion of stability but not in the numerical one. In addition, although it produces the background profile, the planet is assumed to have no effect on the linear stability analysis discussed here, but it is fully incorporated in non-linear simulations. The correspondence between the various approaches adopted indicates the above approximations are reasonable. The discussion of the outer and inner disc sections is very similar, accordingly these are considered below together.

4.6 Angular momentum flux balance for normal modes

The vertically integrated angular momentum flux associated with perturbations can be related to the background vortensity profile. For this discussion, we use the governing equation in the form given by equation (33) (Papaloizou & Savonije 1991). We assume a barotropic disc model and neglect terms involving $\mathrm{d}c_{\mathrm{s}}/\mathrm{d}r$. Multiplying this equation by S'^* and integrating, we get

$$\begin{aligned} \mathcal{F}(S', \sigma) = {} & -\int_{r_{\mathrm{i}}}^{r_{\mathrm{o}}} r\Sigma' S'^* \mathrm{d}r \pm \left[\frac{r\Sigma}{D}\left(\frac{\mathrm{d}S'}{\mathrm{d}r} + \frac{2m\Omega\bar{\sigma}}{r\kappa^2}S'\right)S'^*\right]\Bigg|_{r_{\mathrm{b}}} \\ & - \int_{r_{\mathrm{i}}}^{r_{\mathrm{o}}} \frac{r\Sigma}{D}\left(\frac{\mathrm{d}S'}{\mathrm{d}r} + \frac{2m\Omega\bar{\sigma}}{r\kappa^2}S'\right)\left(\frac{\mathrm{d}S'^*}{\mathrm{d}r} + \frac{2m\Omega\bar{\sigma}}{r\kappa^2}S'^*\right)\mathrm{d}r \\ & - \int_{r_{\mathrm{i}}}^{r_{\mathrm{o}}} \frac{4m^2\Omega^2\Sigma|S'|^2}{r\kappa^4}\mathrm{d}r + \int_{r_{\mathrm{i}}}^{r_{\mathrm{o}}} \frac{m|S'|^2}{\bar{\sigma}}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{\eta}\right)\mathrm{d}r = 0, \end{aligned} \tag{36}$$

where, here and below, the positive sign alternative applies to the case where the disc domain has an inner sharp edge where $r = r_{\mathrm{i}}$ and then $r_{\mathrm{b}} = r_{\mathrm{o}}$, the outer boundary radius. The negative sign alternative applies to the case where the disc domain has an outer sharp edge where $r = r_{\mathrm{o}}$ and then $r_{\mathrm{b}} = r_{\mathrm{i}}$, the inner boundary radius. In both cases waves may propagate through $r = r_{\mathrm{b}}$. Note also that there is no contribution from edge boundary terms as the surface density is assumed to be negligible there.

We may now express the terms in the above integral relation, which for convenience we have taken to define a functional of S' that depends on σ, in terms of quantities related to the transport of angular momentum by taking its imaginary part. We shall begin by assuming that σ is real or, more precisely, that we are in the limit that marginal stability is approached.

By making use of equation (25) we find that

$$\begin{aligned} \mathrm{Im}\left[\int_{r_{\mathrm{i}}}^{r_{\mathrm{o}}} r\Sigma' S'^* \mathrm{d}r\right] &= \mathrm{Im}\left[\int_{r_{\mathrm{i}}}^{r_{\mathrm{o}}} r\Sigma'\left(\frac{\Sigma'^* c_{\mathrm{s}}^2}{\Sigma} + \Phi'^*\right)\mathrm{d}r\right] \\ &= \mathrm{Im}\left[\int_{r_{\mathrm{i}}}^{r_{\mathrm{o}}} r\Sigma'\Phi'^* \mathrm{d}r\right] = \pm\frac{2}{m}\left[rF_{\mathrm{Gr}}\right]|_{r_{\mathrm{b}}}. \end{aligned} \tag{37}$$

We recall that F_{Gr} is the vertically integrated angular momentum flux transported by gravitational stresses. Note that F_{Gr} at the sharp

edge is ignored. Equation (25), together with the requirement that $F_{\rm Gr}$ is regular for $r \to 0$ and vanishes more rapidly than $1/r$ for $r \to \infty$, implies that $F_{\rm Gr}$ is zero at the sharp edge separating the disc domain and the exterior (assumed) vacuum or very low density region.

In addition we remark that from equations (31) and (32), the radial Lagrangian displacement is given by

$$\xi_{\rm r} = \frac{u'_{\rm r}}{{\rm i}\bar{\sigma}} = -\frac{1}{D}\left(\frac{{\rm d}S'}{{\rm d}r} + \frac{2mS'\Omega}{r\bar{\sigma}}\right). \tag{38}$$

From this it follows that the imaginary part of the second term on the right-hand side (RHS) of equation (36) is $-\pm{\rm Im}[r\Sigma\xi_r S'^*]|_{r_{\rm b}} = -\pm(2/m)[rF_{\rm Ar}]|_{r_{\rm b}}$. Using the above, taking the imaginary part of (36) yields the remarkably simple expression

$$\pm\,[r(F_{\rm Gr}+F_{\rm Ar})]\,|_{r_{\rm b}} = \frac{m}{2}{\rm Im}\left[\int_{r_{\rm i}}^{r_{\rm o}} \frac{m|S'|^2}{\bar{\sigma}}\frac{\rm d}{{\rm d}r}\left(\frac{1}{\eta}\right){\rm d}r\right]. \tag{39}$$

Note that we have been assuming that σ is real and that the mode is at marginal stability. Then the RHS of (39) is apparently real. However, the integrand is potentially singular at corotation where $\bar{\sigma} = 0$. Thus the approach to marginal stability has to be taken with care.

Setting $\sigma = \sigma_{\rm R} - {\rm i}\gamma$, where $\sigma_{\rm R}$, being the real part of σ, defines the corotation radius, $r_{\rm c}$, through $\Omega(r_{\rm c}) = -\sigma_{\rm R}/m$ and the growth rate as γ, with $-\gamma$ being the imaginary part of σ. Marginal stability can be approached by assuming that γ has a vanishingly small magnitude but is positive. Then we can use the Landau prescription, and substitute $1/\bar{\sigma}$ by $\mathcal{P}(1/\bar{\sigma}) + {\rm i}\pi\delta(\bar{\sigma})$, where $\mathcal{P}$ indicates that the principal value of the integral is to be taken and δ denotes Dirac's delta function. Adopting this, equation (39) becomes

$$[r(F_{\rm Gr}+F_{\rm Ar})]\,|_{r_{\rm b}} = \pm\frac{\pi|m|}{2}\left[\frac{|S'|^2}{|{\rm d}\Omega/{\rm d}r|}\frac{\rm d}{{\rm d}r}\left(\frac{1}{\eta}\right)\right]\Bigg|_{r_{\rm c}}. \tag{40}$$

The above relation can be viewed as stating that at marginal stability, either corotation is at a vortensity extremum and both terms in equation (40) vanish or angular momentum losses as a result of waves passing through $r = r_{\rm b}$ are balanced by torques exerted at corotation. The latter torque is proportional to the gradient of the vortensity η (Goldreich & Tremaine 1979) and recall that η^{-1} scales with Σ.

The propagating waves quite generally carry negative angular momentum when located in an inner disc and positive angular momentum when located in an outer disc (Goldreich & Tremaine 1979). Thus a balance may be possible between angular momentum losses resulting from the propagation of these waves out of the system and angular momentum gained or lost by an edge disturbance as expressed by equation (40). For an edge disturbance in a region of increasing (decreasing) surface density, such as an inner (outer) edge to an outer (inner) disc, the edge disturbance loses (gains) angular momentum.

In practice, an inner (outer) edge disturbance may be associated with a positive (negative) surface density slope near a vortensity maximum or near a vortensity minimum occurring as a result of the variation of both κ^2 and Σ. The former case is associated with discs for which self-gravity is important, and spiral density waves are readily excited. The latter is associated with weakly or NSG discs and is associated with vortex formation at gap edges as has already been discussed by several authors [see e.g. Lin & Papaloizou (2010), and references therein].

So far we have not discussed the boundary condition at $r = r_{\rm b}$ (the non-sharp boundary). One possibility is that this corresponds to stipulating outgoing waves or a radiation condition, so that the fluxes in (40) are non-zero. Another possibility is to have a surface density taper to zero which would mean wave reflection and removal of the boundary flux terms. In fact, we expect that the issue of stability is not sensitive to this. We remark that in the case of NSG discs and the low m modes of interest, the regions away from the edge are evanescent and amplitudes are small there, the disturbance being localized in the vicinity of the edge. For SG discs, as we indicate below, instability can appear because of an unstable interaction between *outwardly propagating positive* (*inwardly propagating negative*) angular momentum density waves and a *negative* (*positive*) angular momentum edge disturbance localized near an *inner* (*outer*) disc gap edge. As long as these waves are excited it should not matter whether they are reflected or transmitted at large distance, as long as their angular momentum density remains in the wider disc and is not fed back to the exciting edge disturbance.

For these reasons and also for simplicity, we shall assume that at the boundary where $r = r_{\rm b}$, either the surface density has tapered to zero or there is reflection such that the boundary terms in expressions like (36) may be dropped. equation (40) then simply states that marginal stability occurs when corotation is at a vortensity extremum. We find that the vortensity *maximum* case is associated with an instability in SG discs that is associated with strong spiral waves. On the other hand, the vortensity *minimum* case is associated with the vortex-forming instability in NSG discs that has been previously studied [see Lin & Papaloizou (2010), and references therein] and will not be discussed further in this paper [but see Lin & Papaloizou (2011), for a discussion of the effect of weak self-gravity on vortex modes].

5 DESCRIPTION OF THE EDGE DISTURBANCE ASSOCIATED WITH A VORTENSITY MAXIMUM IN A SELF-GRAVITATING DISC

We now focus on the description of the instability in an SG disc as being due to a disturbance associated with the edge causing the excitation of spiral waves that propagate away from it resulting in destabilization. This occurs because the emitted waves carry away angular momentum which has the opposite sign to that associated with the edge disturbance [see Section 4.6 and equation (40)]. Accordingly the excitation process is expected to lead to the growth of this disturbance.

We thus consider the disturbance to be localized in the vicinity of the edge and the potential perturbation it produces to excite spiral density waves in the bulk of the disc. Our numerical calculations show that edge-dominated modes of this type occur for low m and are dominant in the non-linear regime. We assume the mode is weakly growing corresponding to the back reaction of the waves on the edge disturbance being weak. Thus in the first instance we calculate a neutral edge disturbance and then calculate the wave emission as a perturbation. We emphasize again that an important feature is that corotation is at a vortensity *maximum* (in contrast to the NSG case for which corotation is located at a vortensity *minimum*). We show that such modes require self-gravity. In an average sense they require a sufficiently small Q value and so cannot occur in the NSG limit ($Q \to \infty$).

5.1 Neutral edge disturbances with corotation at a vortensity maximum

We start from equation (33) for Σ'. We assume that only the third term on the RHS need be retained. This is because this term should

dominate near corotation in the presence of large vortensity gradients that are presumed to occur near the edge and the disturbance is assumed localized there [see also the discussion of groove modes in collisionless particle discs of Sellwood & Kahn (1991), where similar assumptions are made]. Thus we have

$$\Sigma' = \frac{S'}{r\bar{\omega}}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{\eta}\right), \tag{41}$$

where $\bar{\omega} = \bar{\sigma}/m$. The associated gravitational potential is given by

$$\Phi' = -G\int_{r_\mathrm{i}}^{r_\mathrm{o}} K_m(r,r')\frac{S'(r')}{\bar{\omega}(r')}\frac{\mathrm{d}}{\mathrm{d}r'}\left[\frac{1}{\eta(r')}\right]\mathrm{d}r'. \tag{42}$$

From the relation $S = \Sigma' c_\mathrm{s}^2/\Sigma + \Phi'$, we thus obtain the integral equation:

$$S'(r)\left[1 - \frac{c_\mathrm{s}^2}{r\Sigma\bar{\omega}}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{\eta}\right)\right] = -\int_{r_\mathrm{i}}^{r_\mathrm{o}} GK_m(r,r')\frac{S'(r')}{\bar{\omega}(r')}\frac{\mathrm{d}}{\mathrm{d}r'}\left[\frac{1}{\eta(r')}\right]\mathrm{d}r'. \tag{43}$$

For a disturbance dominated by self-gravity, pressure should be negligible. Under such an approximation, we have $S' \sim \Phi'$. Equation (41) then implies that to obtain a negative (positive) potential perturbation for a positive (negative) surface density perturbation, we need $\mathrm{sgn}(\bar{\omega}) = \mathrm{sgn}(\mathrm{d}\eta/\mathrm{d}r)$. At corotation where $\bar{\omega} = 0$ and there is a vortensity extremum, this requirement becomes $\mathrm{sgn}(\mathrm{d}\bar{\omega}/\mathrm{d}r) = \mathrm{sgn}(\mathrm{d}\Omega/\mathrm{d}r) = \mathrm{sgn}(\mathrm{d}^2\eta/\mathrm{d}r^2)$. Since typical rotation profiles have $\mathrm{d}\Omega/\mathrm{d}r \equiv \Omega' < 0$, the physical requirement that potential and surface density perturbations have opposite signs implies that $\mathrm{d}^2\eta/\mathrm{d}r^2 < 0$ at corotation, i.e. the vortensity is a maximum. Thus our discussion applies to vortensity maxima only. For the class of vortensity profiles for which $\mathrm{sgn}(\mathrm{d}\eta/\mathrm{d}r) = \mathrm{sgn}(\bar{\omega})$, we demonstrate below that the integral equation (43) may be transformed into a Fredholm integral equation with symmetric kernel.

Noting that corotation is located at a vortensity maximum, the requirement above implies the vortensity increases (decreases) interior (exterior) to corotation (we comment that the discussion may also be extended to the case when that is true with $\mathrm{d}^2\eta/\mathrm{d}r^2$ vanishing at corotation). As we expect the disturbance to be localized around corotation, it is reasonable to assume that this holds and that we may contract the integration domain $(r_\mathrm{i}, r_\mathrm{o})$ to exclude regions which do not conform without significantly affecting the problem.

Proceeding in this way we introduce the function $\mathcal{H}$ such that

$$S'(r) = \mathcal{Z}(r)\mathcal{H}(r), \tag{44}$$

where

$$\mathcal{Z}(r) = \left[-\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{\eta}\right)\frac{1}{\bar{\omega}}\right]^{-1/2}\left[1 - \frac{c_\mathrm{s}^2}{r\Sigma\bar{\omega}}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{\eta}\right)\right]^{-1/2}. \tag{45}$$

Defining the new symmetric kernel $\mathcal{R}(r,r')$ as

$$\mathcal{R}(r,r') \equiv GK_m(r,r')\mathcal{Y}(r)\mathcal{Y}(r'), \tag{46}$$

where

$$\mathcal{Y}(r) = \left[-\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{\eta}\right)\frac{1}{\bar{\omega}}\right]^{1/2}\left[1 - \frac{c_\mathrm{s}^2}{r\Sigma\bar{\omega}}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{\eta}\right)\right]^{-1/2}, \tag{47}$$

we obtain the integral equation

$$\mathcal{H} = \int_{r_\mathrm{i}}^{r_\mathrm{o}} \mathcal{R}(r,r')\mathcal{H}(r')\,\mathrm{d}r'. \tag{48}$$

In other words, $\mathcal{H}$ is required to be the solution to

$$\lambda\mathcal{H} = \int_{r_\mathrm{i}}^{r_\mathrm{o}} \mathcal{R}(r,r')\mathcal{H}(r')\,\mathrm{d}r', \tag{49}$$

with $\lambda = 1$. However, equation (49) is just a standard homogeneous Fredholm integral equation of the second kind. Thus the existence of a non-trivial solution ($\mathcal{H} \neq 0$) implies $\mathcal{H}$ is an eigenfunction of the kernel $\mathcal{R}$ with unit eigenvalue.

The kernel $\mathcal{R}$ is positive and symmetric. Accordingly the Fredholm equation above has the property that the maximum eigenvalue $\lambda > 0$ and is the maximum of the quantity Λ given by

$$\Lambda = \frac{\displaystyle\int_{r_\mathrm{i}}^{r_\mathrm{o}}\int_{r_\mathrm{i}}^{r_\mathrm{o}} \mathcal{R}(r,r')\mathcal{H}(r')\mathcal{H}^*(r)\,\mathrm{d}r\mathrm{d}r'}{\displaystyle\int_{r_\mathrm{i}}^{r_\mathrm{o}} |\mathcal{H}(r)|^2\mathrm{d}r} \tag{50}$$

over all $\mathcal{H}$. Thus, if it is possible to show that we must always have $\max(\Lambda) < 1$, then it is not possible to have a non-trivial solution corresponding to a neutral mode. Upon application of the Cauchy–Schwarz inequality (twice), one can deduce

$$\Lambda^2 \leq \int_{r_\mathrm{i}}^{r_\mathrm{o}}\int_{r_\mathrm{i}}^{r_\mathrm{o}} |\mathcal{R}(r,r')|^2\mathrm{d}r\mathrm{d}r' = \Lambda_\mathrm{est}^2 \geq \max(\Lambda^2). \tag{51}$$

Thus the existence of a neutral mode of this type is not possible if Λ_est^2 is less than unity.

To relate the necessary condition for mode existence to physical quantities, we can estimate Λ_est^2. The Poisson kernel $K_m(r, r')$ is largest at $r = r'$. Other factors in the numerator of $\mathcal{R}$ involve the factor $1/\bar{\omega}$. We assume these factors have their largest contribution at corotation where $r = r_\mathrm{c}$ and $\bar{\omega} = 0$. Hence, $\Lambda_\mathrm{est}^2 \sim \int_{r_\mathrm{i}}^{r_\mathrm{o}}\int_{r_\mathrm{i}}^{r_\mathrm{o}} |\mathcal{R}(r_\mathrm{c}, r_\mathrm{c})|^2\mathrm{d}r\mathrm{d}r'$. Assuming the edge region has width of order L_c, taken to be much less than the local radius but not less than the local scaleheight, we estimate

$$\Lambda_\mathrm{est} \sim |\mathcal{R}(r_\mathrm{c}, r_\mathrm{c})|L_\mathrm{c} = GK_m(r_\mathrm{c}, r_\mathrm{c})\left|\frac{1}{\Omega'}\left(\frac{1}{\eta}\right)''\right|L_\mathrm{c}\left|1 - \frac{c_\mathrm{s}^2}{r_\mathrm{c}\Sigma\Omega'}\left(\frac{1}{\eta}\right)''\right|. \tag{52}$$

All quantities are evaluated at corotation and the double prime denotes $\mathrm{d}^2/\mathrm{d}r^2$. Because the edge is thin by assumption, we can further approximate the Poisson kernel by

$$K_m(r,r') \sim \frac{2}{\sqrt{rr'}}K_0\left(\frac{m}{\sqrt{rr'}}\sqrt{|r-r'|^2 + \epsilon_\mathrm{g}^2}\right), \tag{53}$$

where K_0 is the modified Bessel function of the second kind of order zero and recall that ϵ_g is the softening length. If pressure effects are negligible (c_s^2 being small), we obtain

$$\Lambda_\mathrm{est} \sim \frac{2GK_0(m\epsilon_\mathrm{g}/r_\mathrm{c})}{r_\mathrm{c}}\left|\frac{1}{\Omega'}\left(\frac{1}{\eta}\right)''\right|_{r_\mathrm{c}}L_\mathrm{c}. \tag{54}$$

Since $\eta = \kappa^2/2\Omega\Sigma$, the requirement that Λ_est^2 exceeds unity leads us to expect that modes of the type we have been discussing can exist only for sufficiently large surface density scales. We have approximately

$$\Lambda_\mathrm{est} \sim \frac{4G\Sigma K_0(m\epsilon_\mathrm{g}/r_\mathrm{c})}{\Omega^2 L_\mathrm{c}}. \tag{55}$$

Thus taking $L_\mathrm{c} \sim H$ and ϵ_g being a fraction $\epsilon_{\mathrm{g}0}$ of the local scaleheight, we obtain $\Lambda_\mathrm{est} \sim 4\ln[1/(m\epsilon_{\mathrm{g}0}h)]/(\pi Q)$. On account of the logarithmic factor, the condition $\Lambda_\mathrm{est} > 1$ suggests that edge disturbances which lead to instabilities can be present for Q significantly larger than unity, as is confirmed numerically. However, a surface density threshold must be exceeded. Thus such modes associated with vortensity *maxima* do not occur in an NSG disc. Finally we remark that although we applied a model assumption to obtain an integral equation with symmetric kernel in the above analysis, which enables the existence of solutions to be shown, the demonstration

of a surface density threshold does not depend on this. The application of the Cauchy–Schwartz inequality to (50) may be carried out in a similar manner for the integral equation (43) leading to similar conclusions. The fundamental quantity is the vortensity profile; equation (54) indicates that if it is not sufficiently peaked, there can be no mode. Similarly, mode existence becomes less favoured for increasing softening and/or increasing m.

At this point we remark that an analysis of the type discussed above does not work for vortensity minima. An equation similar to (49) could be derived, but in this case the corresponding kernel $\mathcal{R}$ would be negative, implying inconsistent negative eigenvalues λ. This situation results from the surface density perturbation leading to a gravitational potential perturbation with the wrong sign to satisfy the condition (41). Thus we do not expect the type of edge disturbance considered in this section to be associated with vortensity minima. The fact that edge modes associated with vortensity maxima require a threshold value of Q^{-1} to be exceeded separates them from vortex-forming modes which are modes associated with vortensity minima and require Q^{-1} not to be above a threshold in order to be effective. For most disc models considered in this paper, with minimum $Q \leq 2$, vortex modes do not occur. Discs with minimum $Q \gtrsim 2$ are considered in Lin & Papaloizou (2011).

5.1.1 Implications for disc–planet systems

For the locally isothermal disc models adopted in this paper, with a small aspect ratio $h = 0.05$, edge modes could appear more easily as the disc is cold. This is because if the disc thickness sets the length-scale of the edge, equations (52), (54) and (55) indicate that, for fixed Q, lowering sound speed and hence the disc aspect ratio increases $\Lambda_{\rm est}$.

We can apply these equations to gaps opened by a Saturn mass planet, which is considered in linear and non-linear numerical calculations later on. Without instabilities, these gaps deepen with time. There is also vortensity mixing in the co-orbital region as fluid elements pass through shocks and repeatedly execute horseshoe turns close to the planet. In a fixed orbit this reduces the gap surface density and the magnitude of the edge vortensity peaks. In this way the conditions for edge modes become less favourable with time. In this case, edge modes are expected to develop early on during gap formation, if at all. However, this effect may be less pronounced if the orbit evolves because the planet migrates.

For larger planetary masses such as Jupiter, a deeper gap will be opened but stronger shocks are also induced, which may lead to stronger vortensity peaks. Thus, in view of potentially competing effects, the conditions for edge modes as a function of planet migration and planetary mass must be investigated numerically and this will be undertaken in future studies.

5.2 Launching of spiral density waves

Although localized at the gap edge, the edge disturbance perturbs the bulk of the disc through its gravitational potential exciting density waves. This is expected to be through torques exerted at Lindblad resonances (Goldreich & Tremaine 1979). When the disc is exterior (interior) to the edge, the wave excitation can be viewed as occurring at the outer (inner) Lindblad resonance, respectively. These resonances occur where $\sigma/m = -\Omega + \kappa/m$. Here the negative (positive) sign alternative applies to the outer (inner) Lindblad resonance, respectively. The perturbing potential is given by (42), and from now on this potential is given the symbol $\Phi_{\rm edge}$.

The total conserved angular momentum flux associated with the launched waves, when they are assumed to propagate out of the system, is given by Goldreich & Tremaine (1979) as

$$F_{\rm T} = \frac{\pi^2 m r \Sigma}{\beta} \left| \frac{{\rm d}\Phi_{\rm edge}}{{\rm d}r} + \frac{2m\Omega\bar{\sigma}\,\Phi_{\rm edge}}{r\kappa^2} \right|^2, \tag{56}$$

where $\beta = 2\kappa(\mp\kappa' - m\Omega')$. These waves carry positive angular momentum outwards when excited by an inner edge, or equivalently negative angular momentum inwards when excited by an outer edge. Thus they will destabilize negative angular momentum disturbances at an inner disc edge or positive angular momentum disturbances at an outer disc edge that cause their emission. They will lead to an angular momentum flux at the boundary where $r = r_{\rm b}$, given by $(2\pi)\,[r(F_{\rm Gr} + F_{\rm Ar})]\,|_{r_{\rm b}} = F_{\rm T}$

5.3 Spiral density waves and the growth of edge modes

We now investigate the effects of wave losses at the non-sharp boundary as a perturbation. To do this we relax the assumption that the surface density tapers to zero at the boundary where $r = r_{\rm b}$. Instead, we adopt a small value there such that self-gravity is not important for the density waves. Thus, we retain the domain $\mathcal{D}$ for evaluating the gravitational potential to be $(r_{\rm i}, r_{\rm o})$. Then, for real frequencies, all the terms in equation (36) apart from the term inversely proportional to $\bar{\sigma}$ and the boundary term associated with the advective angular flux at $r = r_{\rm b}$ are real. The imaginary part of the latter term was shown to be proportional to the wave angular momentum flux. We assume that the mode is close to marginal stability (subscript 'ms' below) with corotation at a vortensity maximum where $\sigma = \sigma_{\rm r}$, and write equation (36) in the form $\mathcal{F}(S, \sigma) = \mathcal{F}(S, \sigma_{\rm r} + \delta\sigma) = 0$. Assuming $\delta\sigma$ is small, we then expand to first order in $\delta\sigma$, obtaining

$$\left.\frac{\partial \mathcal{F}}{\partial \sigma}\right|_{\rm ms} \delta\sigma + \mathcal{F}(S, \sigma_{\rm r}) \equiv (D_{\rm r} + {\rm i}D_{\rm i})\delta\sigma + \mathcal{F}(S, \sigma_{\rm r}) = 0. \tag{57}$$

Differentiating (36) we find

$$D_{\rm r} = \left(\mathcal{P} \int_{r_{\rm i}}^{r_{\rm o}} \frac{m|S|^2}{\eta^2 \bar{\sigma}^2} \frac{{\rm d}\eta}{{\rm d}r} {\rm d}r \right)\Bigg|_{\sigma=\sigma_{\rm r}} - \frac{\partial}{\partial\sigma} \left(\int_{r_{\rm i}}^{r_{\rm o}} \frac{r\Sigma}{D} \left| \frac{{\rm d}S}{{\rm d}r} + \frac{2m\Omega\bar{\sigma}}{r\kappa^2} S \right|^2 {\rm d}r \right)\Bigg|_{\sigma=\sigma_{\rm r}} \tag{58}$$

and

$$D_{\rm i} = -\frac{\pi|S|^2}{m\Omega'|\Omega'|} \frac{{\rm d}^2}{{\rm d}r^2} \left(\frac{1}{\eta} \right)\Bigg|_{\rm ms}. \tag{59}$$

Setting $\delta\sigma = \delta\sigma_{\rm r} - {\rm i}\gamma$, where γ is the growth rate and taking the imaginary part of (57), noting that the imaginary part of the first two terms on the RHS of equation (36) contributes to the imaginary part of $\mathcal{F}$, giving this as ${\rm Im}(\mathcal{F}) = -\pm(2/m)\,[r(F_{\rm Gr} + F_{\rm Ar})]\,|_{r_{\rm b}} = \mp(1/(m\pi))F_{\rm T}$ with $F_{\rm T}$ given above, we obtain

$$\gamma = -\pm\frac{F_{\rm T}}{m\pi D_{\rm r}} + \delta\sigma_{\rm r}\frac{D_{\rm i}}{D_{\rm r}}. \tag{60}$$

Similarly the real part of (57) gives

$$D_{\rm r}\delta\sigma_{\rm r} = -\gamma D_{\rm i} - {\rm Re}(\mathcal{F}), \tag{61}$$

where ${\rm Re}(\mathcal{F})$ is the real part of $\mathcal{F}$. We suppose that the waves emitted by the edge disturbance result in $\gamma \neq 0$, but that the back reaction does not change the location of the corotation point from that of the marginally stable mode (as indicated by numerical results). Thus, corotation remains at the original vortensity maximum, implying that conditions adjust so that $\delta\sigma_{\rm r} = 0$. We then have

$$\gamma = \mp\frac{F_{\rm T}}{m\pi D_{\rm r}}, \tag{62}$$

where the upper (lower) sign applies to an inner (outer) edge. In this case, instability occurs when wave emission causes negative (positive) angular momentum to be transferred to a negative (positive) angular momentum edge disturbance located at an inner (outer) sharp edge.

According to equation (62), to obtain instability ($\gamma > 0$) for an inner edge disturbance we need $D_{\rm r} < 0$ (since $F_{\rm T} > 0$). For a disturbance concentrated at corotation near an inner sharp edge, with the disc lying beyond, the first term on the RHS of (58) dominates. As corotation is at a vortensity maximum, we expect the contribution to the first integral of $D_{\rm r}$ from the region just beyond corotation, where $\mathrm{d}\eta/\mathrm{d}r < 0$, to be negative and the contribution from the region just interior to corotation, where $\mathrm{d}\eta/\mathrm{d}r > 0$, to be positive. Because the region exterior to an inner edge has higher surface density than that interior to it, and the integrand for the first term in $D_{\rm r}$ is proportional to Σ, we may expect the region exterior to corotation dominates the contribution to $D_{\rm r}$, making it negative and therefore unstable. Indeed, we find $D_{\rm r} < 0$ for the numerical fiducial case presented in Section 6.1. In this case, there is instability due to the reaction of the emitted outwardly propagating positive angular momentum wave on the negative angular momentum inner gap edge disturbance. A corresponding discussion applies to a disc with an outer sharp edge.

We comment that the above considerations depend on the excitation of waves at a Lindblad resonance that were transported with a conserved action or angular momentum flux towards a boundary where they were lost. However, our linear calculations and simulations described below indicate lack of sensitivity to such a boundary condition. This can be understood if it is emphasized that the significant issue is that after emission, the wave action density should not return to the edge disturbance responsible for it. This is possible for example if the waves become trapped in a cavity in the wider disc in which case a radiative boundary would not be needed.

6 LINEAR CALCULATIONS

In this section, we present numerical solutions to the linear normal mode problem where the background SG disc contains a gap, presumed to have been opened by a planet. The basic state is obtained from hydrodynamic simulations by azimuthally averaging the surface density and azimuthal velocity component fields to obtain one-dimensional profiles. The basic state is thus axisymmetric. The radial velocity is assumed to be zero.

We adopt a local isothermal equation of state $P = c_{\rm s}^2(r)\Sigma$ with sound speed $c_{\rm s} = h\sqrt{GM_*/r}$ and $h = 0.05$ for consistency with non-linear simulations. The softening prescription used is that presented in Section 2. As for the analytic discussion, the gravitational potential due to the planet and viscosity is neglected. The linearized equations follow from equations (30)–(33) together with equations (34) and (35). However, rather than solving in terms of the quantity $S' = \Sigma' c_{\rm s}^2/\Sigma + \Phi'$, which was convenient for analytic discussion, we find it more convenient here to work in terms of the relative surface density perturbation, $W = \Sigma'/\Sigma$ for which a single governing equation may be written down explicitly. In terms of this, the velocity perturbations $u'_{\rm r}$, u'_{φ} can be written in the form

$$u'_{\rm r} = -\frac{1}{D}\left[\mathrm{i}\bar{\sigma}\left(c_{\rm s}^2\frac{\mathrm{d}W}{\mathrm{d}r} + \frac{\mathrm{d}\Phi'}{\mathrm{d}r}\right) + \frac{2\mathrm{i}m\Omega}{r}\left(c_{\rm s}^2 W + \Phi'\right)\right] \tag{63}$$

$$u'_{\varphi} = \frac{1}{D}\left[\frac{\kappa^2}{2\Omega}\left(c_{\rm s}^2\frac{\mathrm{d}W}{\mathrm{d}r} + \frac{\mathrm{d}\Phi'}{\mathrm{d}r}\right) + \frac{m\bar{\sigma}}{r}\left(c_{\rm s}^2 W + \Phi'\right)\right]. \tag{64}$$

The gravitational potential perturbation Φ' can be expressed in terms of W through equations (34) and (35), and we note that $\kappa^2 \equiv r^{-3}\mathrm{d}(r^4\Omega^2)/\mathrm{d}r$. The derivatives $\mathrm{d}\Phi'/\mathrm{d}r$ and $\mathrm{d}^2\Phi'/\mathrm{d}r^2$ can be computed by replacing $K_m(r, r')$ with $\partial K_m/\partial r$ and $\partial^2 K_m/\partial r^2$, respectively. Inserting (63) and (64) into the linearized continuity equation yields the governing equation for W:

$$\frac{\mathrm{d}}{\mathrm{d}r}\left[\frac{r\Sigma}{D}\left(c_{\rm s}^2\frac{\mathrm{d}W}{\mathrm{d}r} + \frac{\mathrm{d}\Phi'}{\mathrm{d}r}\right)\right] + \left[\frac{2m}{\bar{\sigma}}\left(\frac{\Sigma\Omega}{D}\right)\frac{\mathrm{d}c_{\rm s}^2}{\mathrm{d}r} - r\Sigma\right]W + \left[\frac{2m}{\bar{\sigma}}\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{\Sigma\Omega}{D}\right) - \frac{m^2\Sigma}{rD}\right]\left(c_{\rm s}^2 W + \Phi'\right) \equiv \mathcal{L}(W) = 0. \tag{65}$$

Note that the term in $\mathrm{d}c_{\rm s}^2/\mathrm{d}r$ has been kept for consistency with simulations used to setup the basic state. We checked that this term has negligible effect on the results obtained below, by solving the linear problem without this term. This is because $c_{\rm s}$ varies on a global scale, whereas the edge disturbance is associated with strong local gradients.

To solve equation (65) we discretized it on an equally spaced grid applied to the domain $r = [1, 10]$. In practice this employed $N_{\rm r} = 1025$ grid points. Equation (65) together with the applied boundary conditions (see below) is then converted into a matrix equation of the form $\sum_{j=1}^{N_{\rm r}} \mathcal{L}_{ij}(\sigma)W_j = 0$, where the matrix $[\mathcal{L}_{ij}(\sigma)]$ gives the discretized linear operator, which is a function of the frequency σ, and W_j is an approximation to W at the jth grid point. This problem cannot be solved for any value of σ. This has to be consistent with the condition that the determinant of the system of linear equations be zero. Thus σ is an eigenvalue although it is not an eigenvalue of $(\mathcal{L}_{ij})$ in the conventional sense.

We proceed by taking $(\mathcal{L}_{ij})$ to be a function of σ. For a specified value of σ we solve the usual eigenvalue problem $\mathcal{L}_{ij}(\sigma)W_j = \mu(\sigma)W_i$ for an eigenvalue, μ, which may also be considered to be a function of σ. The Newton–Raphson method is then used to solve the equation $\mu(\sigma) = 0$. The values of σ so obtained are the required eigenvalues associated with the physical normal modes of the system.

Because simulations suggest the important disturbances are associated with the outer gap edge, it is reasonable to search for values of σ such that corotation lies near the outer gap edge. We checked that the reciprocal of the final matrix condition number is small (at the level of machine precision) in order for results to be accepted. For simplicity, the boundary condition $\mathrm{d}W/\mathrm{d}r = 0$ was applied at $r = r_{\rm i} = 1$ and $r = r_{\rm o} = 10$. As discussed in Section 4.6, modes are found to be driven by the back reaction of emitted density waves on a disturbance located at an outer gap edge, a process expected to be insensitive to boundary conditions. Indeed, we find that our results are insensitive as to whether the above or other boundary conditions are used (see Section 6.5).

6.1 A fiducial case

In order to provide a fiducial case, we study the disc model with a gap, for which $Q_{\rm o} = 1.5$, that is adopted in Section 7. However, the simulations described in Section 7 from which the model was extracted had $\nu = 10^{-5}$, whereas our linear calculations described below are for inviscid discs.

The basic state gap profile is illustrated in Fig. 4 where the surface density Σ, the relative deviation of the angular velocity from the Keplerian value $\Omega/\Omega_{\rm k} - 1$ and the vortensity η are plotted. As expected from our analytic discussion of modes near to marginal stability, local extrema in η in the vicinity of gap edges are closely

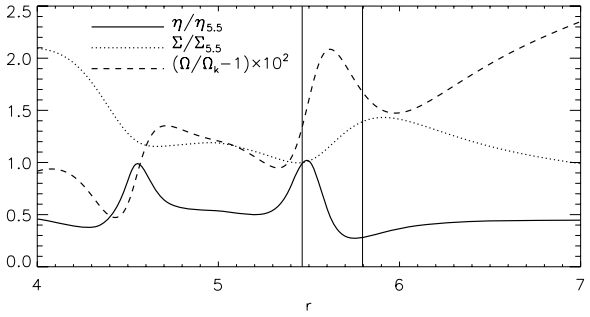

Figure 4. Gap profile produced by a Saturn mass planet in the $Q_o = 1.5$ disc with viscosity $\nu = 10^{-5}$. The surface density Σ and vortensity η are scaled by their values at $r = 5.5$. The relative deviation of the angular velocity from the Keplerian value is also plotted. Vertical lines indicate corotation radii r_c [where $\mathrm{Re}(\bar{\sigma}) = 0$]. For the $m = 2$ mode, $r_c = 5.5$, close to a vortensity maximum. When self-gravity is neglected, $r_c = 5.8$, close to a vortensity minimum.

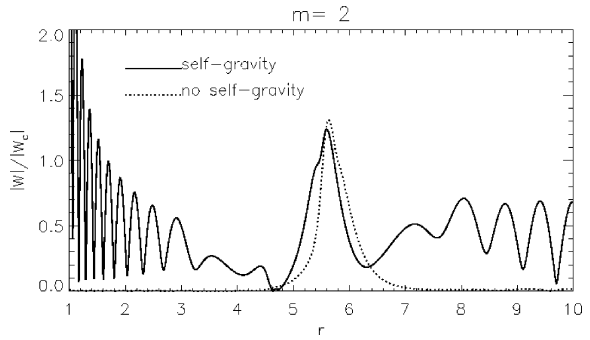

Figure 5. $m = 2$ unstable modes found from linear stability analysis for the fiducial model, with self-gravity (solid) and without self-gravity (dotted). $|W|$ has been scaled by its value at corotation.

associated with instability. This is manifested through mode corotation points being very close to them.

The magnitude of relative surface density perturbation, $|W|$, for SG and NSG responses are shown in Fig. 5. For NSG cases we set $\Phi' = 0$. The eigenfrequencies for SG (NSG) modes are given by $-\sigma = 0.1587 + \mathrm{i}0.4515 \times 10^{-2}$ ($-\sigma = 0.1458 + \mathrm{i}0.2041 \times 10^{-2}$). These correspond to corotation points at $r_c = 5.4626$ and 5.7959 for the SG and NSG modes, respectively. The SG growth rate corresponds to ∼3 times the local orbital period. Thus, although $\gamma/\sigma_R \sim 0.03$ is relatively small, the instability grows on a dynamical time-scale.

The corotation points of SG and NSG modes are very close to local maximum and minimum of η, respectively (Fig. 4). The SG mode grows twice as fast as the NSG mode, consistent with the observation made when comparing edge modes and vortex modes in Section 3, where the former became non-linear sooner.

The SG and NSG eigenfunctions are similar around corotation ($r \in [5, 6]$), although the SG mode has a larger width and is shifted slightly to the left (Fig. 5). As expected from the discussion in Section 5.2, the SG mode has significant wave-like region interior to the inner Lindblad resonance ($r \leq 3.4$) and exterior to the outer Lindblad resonance at ($r \geq 7.2$). By contrast, the NSG mode has negligible amplitude outside [5, 7] compared to that at corotation, whereas the SG amplitude for $r \in [8, 10]$ can be up to $\simeq 56$ per cent of the peak amplitude near corotation. Increasing m in the NSG calculation increases the amplitude in the wave regions, but even for $m = 6$, we found the waves in [8, 10] for the NSG case have an amplitude of about 33 per cent of that at corotation, a smaller value than that pertaining to the SG case with $m = 2$.

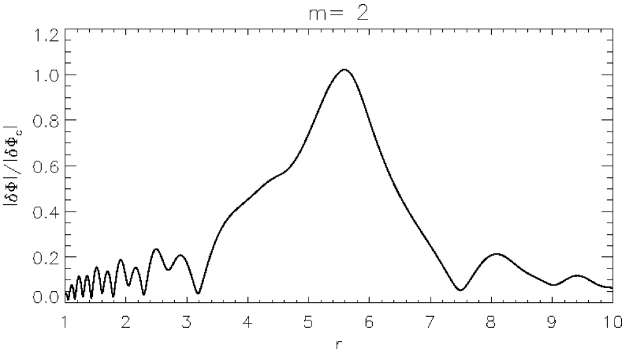

Figure 6. Gravitational potential perturbation Φ' ($\equiv \delta\Phi$ in the plot), scaled by its value at $r = 5.5$ for the $m = 2$ mode with self-gravity.

The behaviour in the wider disc, away from corotation, shows that the for low m the NSG mode is a localized vortex mode, whereas the SG mode corresponds to an edge mode with global spirals. Noting that maxima in the vortensity and Toomre Q nearly coincide, it makes sense that the SG mode is global because away from corotation, the background Toomre Q is decreasing, which makes it easier to excite density waves, because the evanescent zones between corotation and the Lindblad resonances that are expected from WKBJ theory narrow accordingly.

Comparing the SG and NSG modes shown in Fig. 5, we see that including self-gravity in the linear response enables additional waves in the disc at low m. Fig. 6 shows the gravitational potential perturbation for the SG mode. A comparison with $|W|$ in Fig. 5 shows that the surface density perturbation around r_c has an associated potential perturbation that varies on a more global scale. The peak in $|W|$ about r_c is confined to $r \in [4.6, 6.2]$, whereas that for $|\Phi'|$ is confined to $r \in [3.2, 7.5]$, overlapping Lindblad resonances (see Fig. 2) in the latter case. Thus Φ' is less localized around corotation than $|W|$.

Rapid oscillations seen in $|W|$ for $r > 7$ are not observed for $|\Phi'|$ in the same region. Furthermore, $\Phi'(r > 7)$ is at most $\simeq 20$ per cent of the corotation amplitude, which is a smaller ratio than that for $|W|$. This leads to the notion that the disturbance at the outer gap edge is driving the disturbance in the outer disc as in our analytical discussion given above. In effect, the disturbance at corotation acts like an external perturber (e.g. a planet) to drive density waves in the outer disc, through its gravitational field. Clearly, this is only possible in an SG disc.

6.2 Energy balance of edge and wave-like disturbances

The analysis in Section 5 describes the edge mode as being associated with a coupling between disturbances associated with vortensity extrema near an edge and the smooth regions interior or exterior to the edge. We apply this idea to the reference case (with self-gravity) by considering the region $r > r_p$. Specifically, we focus on $r \in [5, 10]$ because hydrodynamic simulations indicate the spiral arms are more prominent in the outer disc (Sections 3 and 7).

Multiplying the governing equation (65) by S'^*, integrating over $[r_1, r_2]$ and then taking the real part, we obtain

$$\mathrm{Re}\int_{r_1}^{r_2} \varrho \mathrm{d}r = \mathrm{Re}\int_{r_1}^{r_2} (\varrho_{\mathrm{corot}} + \varrho_{\mathrm{wave}})\,\mathrm{d}r, \tag{66}$$

where

$$\varrho \equiv r\Sigma W S'^{*} = r\left(c_{\rm s}^2|\Sigma'|^2/\Sigma + \Sigma'\Phi'^{*}\right), \tag{67}$$

$$\varrho_{\rm corot} \equiv \frac{2m}{\bar{\sigma}}\frac{\rm d}{{\rm d}r}\left(\frac{\Sigma\Omega}{D}\right)|S'|^2, \tag{68}$$

$$\varrho_{\rm wave} \equiv S'^{*}\frac{\rm d}{{\rm d}r}\left[\frac{r\Sigma}{D}\left(c_{\rm s}^2\frac{{\rm d}W}{{\rm d}r} + \frac{{\rm d}\Phi'}{{\rm d}r}\right)\right] + \frac{2m}{\bar{\sigma}}\left(\frac{\Sigma\Omega}{D}\right)\frac{{\rm d}c_{\rm s}^2}{{\rm d}r}WS'^{*} - \frac{m^2\Sigma}{rD}|S'|^2. \tag{69}$$

ϱ has dimensions of energy per unit length and, for a normal mode, its real part is four times the energy per unit length associated with pressure perturbations (the term $\propto c_{\rm s}^2$) and gravitational potential perturbations (the term $\propto\Phi'*$). As the factor of four is immaterial to the discussion, we simply call the integral of this quantity over the region concerned the thermal-gravitational energy (TGE). We remark that when self-gravity dominates, the TGE is negative and when pressure dominates, it is positive.

When the integral is performed, one sees that the TGE is balanced by various terms on the RHS of equation (66). For simplicity, we split the terms on the RHS into just two parts that are integrals of $\varrho_{\rm corot}$ and $\varrho_{\rm wave}$ over the region of interest. This is of course not a unique procedure. The vortensity term $\varrho_{\rm corot}$ has been isolated because it contains the potential corotation singularity which can be amplified by large vortensity gradients at the gap edge. The rest of the RHS is collected into $\varrho_{\rm wave}$. Note that the term in ${\rm d}c_{\rm s}^2/{\rm d}r$ is included in $\varrho_{\rm wave}$, despite being proportional to $1/\bar{\sigma}$. This is motivated by trying to keep $\varrho_{\rm corot}$ as close to the vortensity term identified in the analytical formalism as possible. However, we have considered attributing the ${\rm d}c_{\rm s}^2/{\rm d}r$ term to $\varrho_{\rm corot}$ or neglecting it altogether. In both cases, it made negligible difference compared to the splitting adopted above. Again, this is because $c_{\rm s}^2$ varies on a much larger scale than vortensity gradients.

It is important to note that while the real part of the *combination* $\varrho_{\rm corot} + \varrho_{\rm wave}$ gives rise to the TGE, we cannot interpret ${\rm Re}(\varrho_{\rm corot})$ as an energy density of the corotation region. It contains a term contributing to the TGE that is proportional to the vortensity gradient (see below) and potentially associated with a corotation singularity. Similar arguments apply to $\varrho_{\rm wave}$ which, in the strictly isothermal case, can be seen to be associated with density waves (see Section 4.6 above). We use this splitting to show that for the modes of interest, the vortensity term contributes most to the TGE.

Numerically, however, quantities defined above are inconvenient because of the vanishing of D for neutral modes at Lindblad resonances. To circumvent this we work with $D^2\varrho$, $D^2\varrho_{\rm corot}$, and $D^2\varrho_{\rm wave}$. We call these modified energy densities. This change does not disrupt our purpose because the most important balance turns out to be between the terms involving ${\rm Re}(D^2\varrho)$ and ${\rm Re}(D^2\rho_{\rm corot})$, both focused near corotation, where $D \sim \kappa^2$. In the region near Lindblad resonances and beyond, ${\rm Re}(\varrho)$ and ${\rm Re}(\varrho_{\rm corot})$ are small. Thus the incorporation of the D^2 factor does not influence conclusions about the TGE balance. The modified energy densities are plotted in Fig. 7 for the $m = 2$ mode in the fiducial case. The curves share essential features with those obtained using the original definitions without the additional factor of D^2 (these are presented and discussed in Appendix A).

Fig. 7 shows that ${\rm Re}(D^2\varrho)$ is negative around corotation which is located near the outer gap edge, $r \in [5, 6]$). Accordingly ${\rm Re}(\varrho) < 0$ and must be dominated by the gravitational energy contribution since the pressure contribution is positive definite. This supports

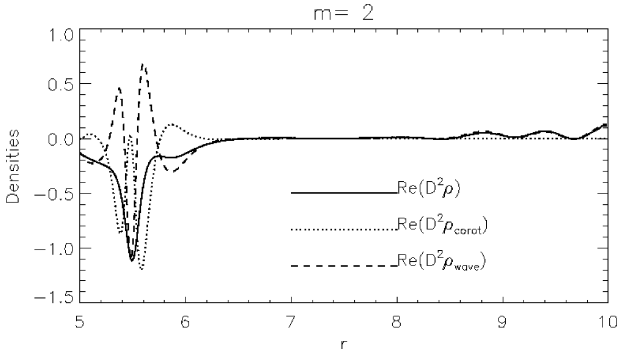

Figure 7. One-dimensional modified energy densities computed from the eigenfunctions W, Φ' for the fiducial case $Q_{\rm o} = 1.5$.

the interpretation of the disturbance as an edge mode where self-gravity is important. If it were the vortex mode then ${\rm Re}(D^2\varrho) > 0$ near corotation, because in that case self-gravity is unimportant compared to pressure.

For $r \geq 8.4$, ${\rm Re}(D^2\rho)$ becomes positive due to pressure effects and oscillates towards the outer boundary as the perturbation becomes wave-like. Similarly, the pressure perturbation dominates towards the inner boundary (not shown). This signifies that self-gravity becomes unimportant relative to pressure within these regions. This is consistent with the behaviour of the gravitational potential perturbations, which are largest near corotation. We checked explicitly that the gravitational energy contribution to the TGE is largely focused around the gap edge ($r \in [5, 6]$), even more so than the gravitational potential perturbation.

Fig. 7 shows that ${\rm Re}(D^2\rho_{\rm corot})$ and ${\rm Re}(D^2\rho_{\rm wave})$ have their largest amplitudes for $r \in [5, 6]$. Integrating over $r \in [5, 10]$, we find $\tilde{U} < 0$, and using a normalization such that

$$\tilde{U} \equiv {\rm Re}\int_5^{10} D^2\varrho{\rm d}r = -|\tilde{U}|,$$

we find

$$\tilde{U}_{\rm corot} \equiv {\rm Re}\int_5^{10} D^2\varrho_{\rm corot}{\rm d}r \simeq -0.822|\tilde{U}|,$$

$$\tilde{U}_{\rm wave} \equiv {\rm Re}\int_5^{10} D^2\varrho_{\rm wave}{\rm d}r \simeq -0.179|\tilde{U}|.$$

This implies the TGE is negative, and hence a gravitationally dominated disturbance. We have $|\tilde{U}_{\rm corot}/\tilde{U}| \sim 0.8$, suggesting the TGE is predominantly balanced by the vortensity term which from Fig. 7, is localized in the gap edge region, balances the gravitational energy of the mode. Because vortensity gradients are largest near the gap edge, gravitational energy is most negative here overtaking pressure, resulting in a negative TGE.

6.2.1 *Further analysis of the vortensity term*

The vortensity term $\rho_{\rm corot}$ that we defined above differs from that naturally identified in the alternative form of the governing equation in Section 4. However, this difference is not significant. The vortensity term can be further decomposed as

$$\varrho_{\rm corot} = \varrho_{\rm corot,1} + \varrho_{\rm corot,2}$$

$$\varrho_{\rm corot,1} = \frac{m}{\bar{\sigma}(1-\bar{\nu}^2)}\frac{\rm d}{{\rm d}r}\left(\frac{1}{\eta}\right)|S'|^2,$$

$$\varrho_{\rm corot,2} = \frac{2m}{\kappa\eta(1-\bar{\nu}^2)^2}\frac{{\rm d}\bar{\nu}}{{\rm d}r}|S'|^2,$$

where $\bar{\nu} = \bar{\sigma}/\kappa$. Let us now temporarily consider $\varrho_{\rm corot,1}$ as the new vortensity term, so that $\varrho_{\rm corot} \rightarrow \varrho_{\rm corot,1}$ and attribute $\varrho_{\rm corot,2}$ to the wave term so that $\varrho_{\rm wave} \rightarrow \varrho_{\rm wave} + \varrho_{\rm corot,2}$. The new vortensity term is proportional to the vortensity gradient explicitly. With the new definitions, we find

$$|\tilde{U}_{\rm corot}/\tilde{U}| \sim 0.690.$$

Thus the new corotation term alone accounts for ∼70 per cent of the (modified) TGE and giving almost the same result as the previous one. In addition, as most of the contribution comes from around corotation, where $|\bar{\nu}| \ll 1$. Replacing the factor $(1-\bar{\nu}^2)$ by unity has a negligible effect on the results. However, making this replacement gives

$$\rho_{\rm corot,1} \rightarrow \frac{m|S'|^2}{\bar{\sigma}}\frac{\rm d}{{\rm d}r}\left(\frac{1}{\eta}\right),$$

which corresponds to the vortensity term used in the analysis given in Section 4. Whether this term contributes positively or negatively to the TGE depends on the sign of $\int_{r_1}^{r_2}\varrho_{\rm corot,1}{\rm d}r$. We expect most of the contribution to this integral comes from corotation where $\bar{\sigma} \simeq 0$. Then ${\rm sgn}(\int_{r_1}^{r_2}\varrho_{\rm corot,1}{\rm d}r) = {\rm sgn}(\varrho_{\rm corot,1}|_{r_{\rm c}})$. Evaluating this for a neutral mode with corotation at a vortensity extremum, we find $\varrho_{\rm corot,1}|_{r_{\rm c}} = -\eta^{-2}|S'|^2({\rm d}^2\eta/{\rm d}r^2)/\Omega'|_{r_{\rm c}}$. Since quite generally, $\Omega' < 0$, we conclude that if corotation occurs at a maximum value of η, then the vortensity term contributes negatively to the TGE. As stated above, when self-gravity dominates pressure, the TGE is negative, so it can then be mostly accounted for by the vortensity term. This is precisely the type of balance assumed for the neutral edge mode modelled in Section 5.1.

However, a consequence of the above discussion is that if corotation occurs at a minimum value of η, then the vortensity term contributes positively to the TGE. This can only give the dominant contribution when the TGE is positive, which can only be the case when the pressure contribution dominates over that from self-gravity. In the limit of weak self-gravity, the TGE will be >0 and it can be balanced by a localized disturbance with corotation at a vortensity minimum as happens for the vortex modes.

6.3 Dependence on *m* and disc mass

We have solved the linear eigenmode problem for discs with $1.2 \leq Q_{\rm o} \leq 1.6$. The basic state is set up from hydrodynamic simulations as described in Section 3, with $\nu = 10^{-5}$. The local max(Q) near the outer gap edge ranges between $Q = 3.3$ and 4.4. Self-gravity is included in the response ($\Phi' \neq 0$).

Fig. 8(a) compares eigenfunctions , $|W|$, for different m for the $Q_{\rm o} = 1.2$ disc. Increasing m increases the amplitude in the wave region relative to that around corotation ($r_{\rm c} \simeq 5.5$), resulting in $m = 6$ being more global than $m = 3$. High m modes do not fit our description of edge modes in the region of interest ($r > 5$), because the disturbance at corotation becomes comparable to or even smaller than that in the wave region beyond the outer Lindblad resonance. It is then questionable to interpret the waves as a secondary phenomenon induced by the edge disturbance, even though the physics of the modes, as being entities driven by an unstable interaction between edge and wave disturbances, may be more or less the same. Hence, we typically find what we describe as edge-dominated modes with $m \lesssim 3$ in simulations where the surface density perturbation is maximal near the gap edge. For fixed $m = 2$, hence in that regime, Fig. 8(b) shows that lowering $Q_{\rm o}$ makes

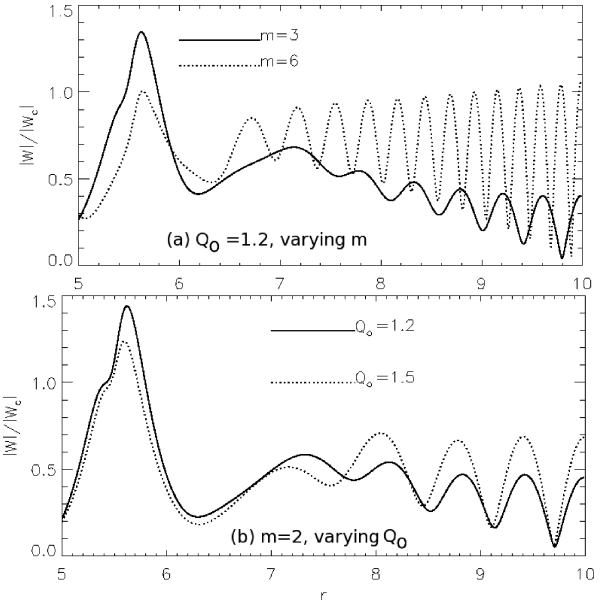

Figure 8. The effect of azimuthal wavenumber m (a) and the effect of disc mass, parametrized by $Q_{\rm o}$ (b) on linear edge modes.

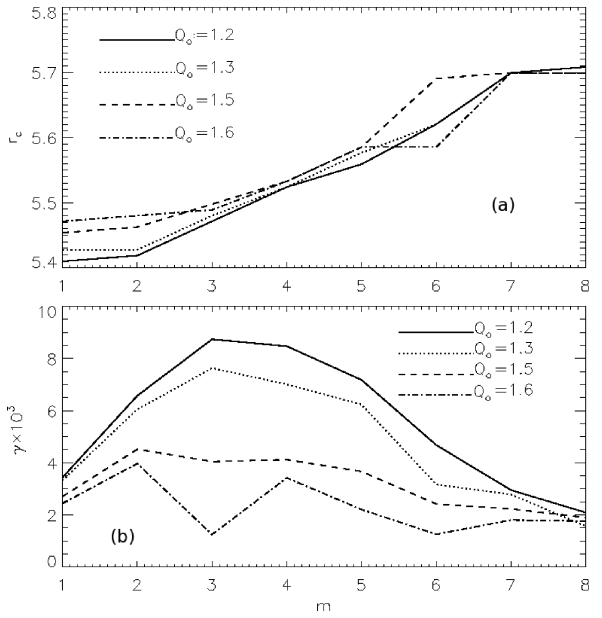

Figure 9. Corotation radii (a) and growth rates (b) as a function of azimuthal wavenumber m, for a range of $Q_{\rm o}$ values.

the corotation amplitude larger relative to that in the wave region. This is expected because increasing the level of self-gravity means the necessary condition for edge disturbance is more easily satisfied (Section 5.1).

In Fig. 9 we plot the locations of the corotation points and the growth rates for the unstable modes we found as a function of the azimuthal mode number, m, and a range of disc masses (parametrized by $Q_{\rm o}$). Note that changing $Q_{\rm o}$ affects both the background state and the linear response, while m acts as a parameter of the linear response only. The plots in Fig. 9(a) show that the corotation radius tends to move outwards with increasing m and/or $Q_{\rm o}$. However, corotation is always located in the edge region where the vortensity

is either decreasing or maximal. The largest growth rates are found for low m ($\lesssim 3$) edge-dominated modes.

Referring to the basic state (Fig. 4), we see that increasing m and/or Q_o, which weakens self-gravity, shifts corotation towards $\max(\eta^{-1})$ (the vortensity minimum), which disfavours edge-dominated modes. For sufficiently low mass or NSG discs, we expect corotation associated with vortensity minima and therefore vortex modes to dominate (e.g. $Q_o = 4$ in Section 3). Thus we only expect edge-dominated modes for low m and sufficiently large disc mass. An exception to the general trend above is that the $Q_o = 1.6, m = 6$ mode has a smaller corotation radius than the corresponding mode with $Q_o < 1.6$. This may be a boundary effect because for this value of m and Q_o the modes are distributed through the outer disc and require the implementation of accurate radiative boundary conditions, rather than the simplified boundary conditions actually applied. However, such high m modes typically have growth rates smaller than those for lower m, accordingly we do not expect them to dominate, nor are they observed in the non-linear simulations discussed later.

Fig. 9(b) shows that growth rates increase as Q_o is lowered (increasing surface density scale). We found that decreasing Q_o results in deeper gaps, because disc material trapped in the vicinity of the planet adds to the planet potential, so that the effective planet mass is larger for smaller Q_o. Steeper gaps are expected to be more unstable, therefore there is a contribution to the increased growth rate from the changes to the background profile as the disc mass is increased.

We also expect the effect of lowering Q_o, through the linear response, to destabilize edge-dominated modes as they rely on self-gravity. However, the effect of self-gravity through the response is weaker for larger m values because the size of the Poisson kernel decreases. Hence, the increase of the growth rates with disc mass for $m = 4$–6 is likely more attributable to the effect of self-gravity on the basic state. Growth rates for the most unstable mode approximately double as the disc mass is increased by 25 per cent ($Q_o = 1.5 \rightarrow Q_o = 1.2$). Note for fixed $Q_o \leq 1.5$, the $m = 3$–5 growth rates are similar but $m > 3$ are not edge modes.

For $Q_o = 1.2, 1.3$, the most unstable mode has $m = 3$, whereas for $Q_o = 1.5$, the $m = 2$ mode has the largest growth rate. As discussed in Section 5, edge-dominated modes require sufficient disc mass and/or low m. For fixed m, they are stabilized with increasing Q_o. Hence, for edge modes to exist with decreasing disc mass, they must shift to smaller m. The higher m modes are less affected by self-gravity. This may explain the double peak in growth rate plotted as a function of m for $Q_o = 1.6$ as we move from $Q_o = 1.5$ (we remark that for $Q_o = 1.5$ the $m = 3$ growth rate is also slightly smaller than $m = 2, 4$).

The integrals of the modified energy densities for each azimuthal wavenumber are given for the fiducial case with $Q_o = 1.5$ in Table 1. Only the modes with $m \leq 2$ are edge-dominated modes because for these sum of the integrals contributing to the (modified) TGE is negative, corresponding to a gravitationally dominated disturbance and over 50 per cent of the energy is accounted by the vortensity edge term. The mode with $m = 2$ had the largest growth rate and corotation radius at $r = 5.46$. This is fully consistent with the non-linear simulation of the fiducial case performed in Section 3.1. There the dominant mode in the outer disc had $m = 2$ and corotation radius at $r \sim 5.5$.

For $m > 2$, the total (modified) energy becomes positive, which must be due to the pressure term $(c_s^2|\Sigma'|/\Sigma)$ becoming dominant. The vortensity term alone cannot balance this because it contributes negatively. For $m \geq 4$, the vortensity term contribution is small consistent with high m solutions being predominantly wave-like (Fig. 8). However, note that in spite of this the vortensity term may still play some role in driving the modes through corotation torques. These trends have also been found for other models so they appear to be general. However, high m global modes are unimportant for the problems we consider, as they are not seen to develop in non-linear simulations.

Table 1. Eigenfrequencies expressed in units of $\Omega_k(r = 1)$, corotation radii and modified energy integrals for different azimuthal mode numbers, m, for the fiducial disc with $Q_o = 1.5$. The modified energy integrals are taken over the outer disc $r \in [5, 10]$.

m	$-\sigma_R$	$\gamma \times 10^2$	r_c	$-\tilde{U}_{corot}/\lvert\tilde{U}\rvert$	$-\tilde{U}_{wave}/\lvert\tilde{U}\rvert$
1	0.079	0.270	5.454	0.65	0.35
2	0.159	0.452	5.463	0.82	0.18
3	0.237	0.404	5.498	0.34	−1.24
4	0.313	0.412	5.533	0.026	−1.03
5	0.386	0.366	5.585	0.0048	−1.004
6	0.462	0.242	5.603	0.0015	−0.998
7	0.524	0.223	5.699	0.0010	−0.997
8	0.599	0.189	5.699	0.00099	−0.995

6.4 Softening length

Gravitational softening has to be used in a two-dimensional calculation. It prevents a singularity at $r = r'$ in the Poisson kernel and approximately accounts for the disc's vertical dimension but is associated with some uncertainty. Apart from the linear response, softening also has an effect through the gap profile set up by our non-linear simulations. A fully self-consistent treatment requires a new simulation with each new softening considered. For reasons of numerical tractability though, we only performed experiments using two values of softening parameter $\epsilon_{g0,e}$ in simulations to set up the gap profile and range of values $\epsilon_{g0,l}$ used in the linear response.

Note that in non-linear simulations a single disc potential softening parameter ϵ_{g0} is used. Our results are summarized in Table 2. The integrated total modified energy values indicate we have found dominated edge modes since the major contribution is due to the corotation/vortensity term.

Table 2. As for Table 1 but for linear calculations with different softening parameters, for the $Q_o = 1.5$ fiducial case and azimuthal mode number $m = 2$. Subscripts 'l' denote the softening parameter used in the linear response, and 'e' denotes the softening parameter used in setting up the basic state. Note that although growth rates vary by about a factor of 2, the location of the corotation radius does not vary significantly.

$\epsilon_{g0,l}$	$\epsilon_{g0,e}$	$-\sigma_R$	$\gamma \times 10^2$	$-\tilde{U}_{corot}/\lvert\tilde{U}\rvert$	$-\tilde{U}_{wave}/\lvert\tilde{U}\rvert$
0.03	0.3	0.159	0.560	0.66	0.33
0.05	0.3	0.159	0.544	0.67	0.32
0.1	0.3	0.159	0.508	0.69	0.30
0.2	0.3	0.159	0.461	0.74	0.23
0.3	0.3	0.159	0.452	0.82	0.18
0.5	0.3	0.158	0.435	0.87	0.14
0.3	0.6	0.159	0.420	0.91	0.09
0.5	0.6	0.158	0.410	0.95	0.064
0.6	0.6	0.158	0.375	0.99	0.0017
0.8	0.6	0.158	0.276	1.15	−0.14

Comparing growth rates for fully self-consistent cases $\epsilon_{g0,e} = \epsilon_{g0,l} = 0.3,\ 0.6$ shows that increasing the softening length stabilizes edge modes, which is expected since they are driven by self-gravity. In addition, all growth rates for $\epsilon_{g0,e} = 0.6$ are smaller than those for $\epsilon_{g0,e} = 0.3$ independently of $\epsilon_{g0,l}$. This indicates stabilization of edge modes via the basic state when softening is increased. For fixed $\epsilon_{g0,e}$, γ decreases as $\epsilon_{g0,l}$ increases. We could not find edge modes for $\epsilon_{g0,l} \geq 0.6$ with $\epsilon_{g0,e} = 0.3$, nor for $\epsilon_{g0,l} \gtrsim 0.83$ with $\epsilon_{g0,e} = 0.6$. An upper limit is expected from analytical considerations (see Section 5.1) because if self-gravity in the linear response is too weak, the vortensity/corotation disturbance cannot be maintained so edge-dominated modes are suppressed. This is intuitive because the edge mode requires gravitational interaction between gap edge and the smooth disc.

6.4.1 Convergence issues

For fixed $\epsilon_{g0,e} = 0.3$, the energy ratios in Table 2 indicate convergence as $\epsilon_{g0,l} \to 0$. Perhaps surprisingly, $|\tilde{U}_{\rm corot}/\tilde{U}|$ decreases with softening. We found this is because $\varrho_{\rm wave}$, as defined by equation (69), includes a term proportional to $\mathrm{d}^2\Phi'/\mathrm{d}r^2$ which is the dominant contribution to $|\tilde{U}_{\rm wave}|$, and its contribution mostly comes from the corotation region (as the potential perturbation is largest there). As the strength of self-gravity is increased by decreasing softening, the edge disturbance is modified, but the subsequent effect of the $\mathrm{d}^2\Phi'/\mathrm{d}r^2$ term on the energies cannot be anticipated a priori.

For $\epsilon_{g0,l} = 0.03$, the vortensity term does not dominate the modified total energy as much. However, this case is in fact not self-consistent because the softening used to set up the gap profile is an order of magnitude larger than that used in linear calculations. The limit of zero softening is also physically irrelevant because the disc has finite thickness. For the self-consistent cases with reasonable softening values $(0.3, 0.6)$, the vortensity term dominates the energy balance, so the analytic description developed above works reasonably well.

6.5 Edge mode boundary issues

The boundary condition $W' = 0$ was applied for simplicity. Vortex modes in low mass or NSG discs are localized and insensitive to boundary conditions (de Val-Borro et al. 2007). The edge-dominated modes rely on self-gravity and are therefore intrinsically global, boundary effects cannot be assumed unimportant a priori.

We performed additional experiments with varying boundary conditions for the fiducial case ($Q_{\rm o} = 1.5$). These include relocating the inner boundary to $r = 1.1, 2.5$ or the outer boundary to $r = 9.8$ and approximating/extrapolating boundary derivatives using interior grid points (Adams, Ruden & Shu 1989). In the last case, the linear ODE is applied at the end points of the grid and therefore no boundary conditions imposed. For SG disc calculations giving edge-dominated modes, these various conditions gave corotation radii varying by about 0.2 per cent and growth rates varying by at most 10 per cent. Overall the eigenfunctions W are similar, showing the essential features of the edge mode, including relative amplitude of outer disc wave region and the corotation region. We conclude that the existence of edge-dominated modes is not too sensitive to boundary effects. Physically this is because edge modes are mainly driven by the local vortensity maximum or edge, which is an internal feature away from boundaries.

7 NON-LINEAR HYDRODYNAMIC SIMULATIONS

We present non-linear hydrodynamic simulations of disc–planet models with $1.2 \leq Q_{\rm o} \leq 2$, corresponding to disc masses $0.079M_* \geq M_{\rm d} \geq 0.047M_*$. Most simulations use a viscosity of $\nu = 10^{-5}$ or equivalently $\alpha = O(10^{-3})$, which has been typically adopted for protoplanetary discs. This value has also been found to suppress vortex instabilities (de Val-Borro et al. 2007). We adopt gravitational softening parameter $\epsilon_{g0} = 0.3$. The planet is set on a fixed circular orbit at $r_{\rm p} = 5$ in order to focus on gap stability. We briefly explore the effects of varying viscosity and softening lengths later in this section and planetary migration in the next section.

7.1 Numerical method

The hydrodynamic equations are evolved with the FARGO code (Masset 2000). FARGO is an explicit finite-difference code similar to ZEUS (Stone & Norman 1992) but customized for disc–planet interactions. It circumvents the time-step limit imposed by the rotational velocity at the inner boundary by splitting the azimuthal velocity into mean and perturbed parts, azimuthal transport being performed on each separately. Self-gravity for FARGO was implemented and tested by Baruteau & Masset (2008). Two-dimensional self-gravity can be calculated using Fast Fourier Transform (Binney & Tremaine 1987). This requires the radial domain to be logarithmically spaced and doubled in extent. The planet potential is introduced at $25P_0$ and its gravitational potential ramped up over $10P_0$, where P_0 is the Keplerian period at $r = 5$.

The disc is divided into $N_{\rm r} \times N_\varphi = 768 \times 2304$ grid points in radius and azimuth giving a resolution of $\Delta r/H = 16.7$. The grid cells are nearly square, with $\Delta r/r\Delta\phi = 1.1$. We impose open boundaries at $r_{\rm i}$ and non-reflecting boundaries at $r_{\rm o}$ (Godon 1996). The latter has also been used in the SG disc–planet simulations of Zhang et al. (2008). As argued above, because edge modes are physically driven by an internal edge, we do not expect boundary conditions to significantly affect whether or not they exist. Indeed, simulations with open outer boundaries or damping boundaries (de Val-Borro 2006) all show development of edge modes.

7.2 Overview

We first present an overview of the effect of edge-dominated modes on disc–planet systems. Fig. 10 shows the relative surface density perturbation ($\Delta\Sigma/\Sigma$) for $Q_{\rm o} = 1.2, 1.5$ and 2.0. The profile formed in these cases has local $\max(Q) = 3.3, 4.2, 5.3$ near the outer gap edge and the average Q for $r \in [6, r_{\rm o}]$ is $1.6, 2.0, 2.6$. Although the edge mode is associated with the local vortensity maximum [located close to $\max(Q)$ for gap profiles], it requires coupling to the external smooth disc via self-gravity. Hence, edge modes only develop if Q in these smooth regions is sufficiently small, otherwise global disturbances cannot be constructed.

The case $Q_{\rm o} = 2.0$ gives a standard result for gap-opening planets, typically requiring planetary masses comparable or above that of Saturn. For that mass a partial gap of depth $\sim$50 per cent is formed, a steady state attained and remains stable to the end of the simulation. This serves as a stable case for comparison with more massive discs. This simulation was repeated with $\nu = 10^{-6}$, in which case we identified both vortex and edge modes. The standard viscosity $\nu = 10^{-5}$ thus suppresses both types of instabilities for $Q_{\rm o} = 2$. Although we adopted $\nu = 10^{-5}$ to avoid complications from vortex modes, we

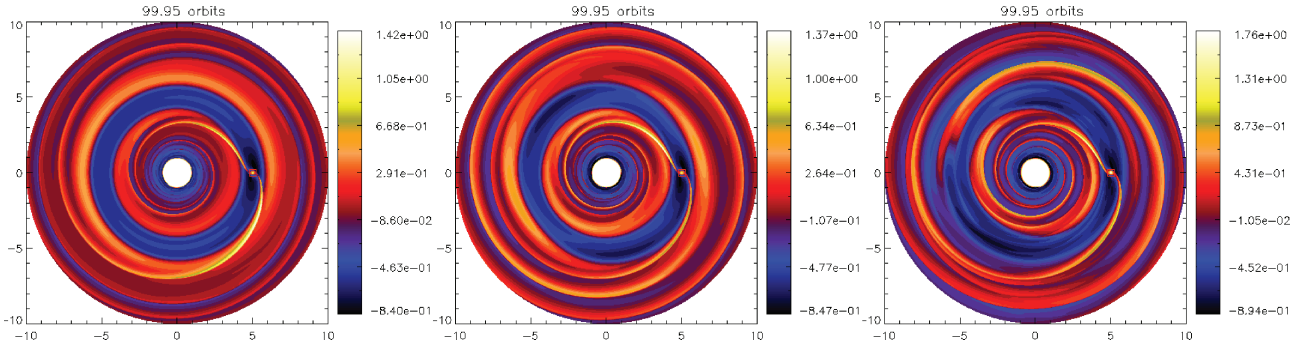

Figure 10. Surface density perturbations (relative to $t = 0$) at $t \sim 100P_0$ as a function the minimum value of Toomre Q in the disc, Q_o. From left to right: $Q_o = 2.0, 1.5, 1.2$.

note that they are stabilized for sufficiently massive discs anyway (e.g. $Q_o = 1.5, \nu = 10^{-6}$ do not show vortices; see Section 3).

Gap edges become unstable for $Q_m \leq 1.5$ when $\nu = 10^{-5}$. As implied by linear analysis, the $m = 2$ edge mode dominates when $Q_o = 1.5$ and it saturates by $t = 100P_0$. It is more prominent in the outer disc, and overdensities deform the gap edge into an eccentric ring. The underdensity in the gap also becomes more non-axisymmetric, being deeper where the gap is wider. The inner disc remains fairly circular, though non-axisymmetric disturbance ($m = 3$) can be identified. Note the global nature of the edge modes in the outer disc: spiral arms can be traced back to disturbances at the gap edge rather than the planet. Consider the spiral disturbance at the outer gap edge just upstream of the planet where $\Delta\Sigma/\Sigma$ is locally maximum. Moving along this spiral outwards, $\Delta\Sigma/\Sigma$ reaches a local minimum at $(x, y) = (6, -4)$ before increasing again. This is suggestive of the outer disc spirals being induced by the edge disturbance.

When the effect of disc self-gravity is increased to the $Q_o = 1.2$ case, there is significant disruption to the outer gap edge, it is no longer clearly identifiable. The $m = 1$ edge mode becomes dominant in the outer disc, although overall it still appears to carry an $m = 2$ disturbance because of the perturbation due to the planet. The inner disc becomes visibly eccentric with an $m = 2$ edge mode disturbance. The shocks due to edge modes can comparable strength to those induced by the planet.

7.3 Development of edge modes for $Q_o = 1.5$

We examine the fiducial case with $Q_o = 1.5$. Fig. 11 shows the development of the edge instability. The final dominance of $m = 2$ is consistent with linear calculations. The planet opens a well-defined gap by $t \sim 40P_0$ or $\sim 5P_0$ after the planet potential , Φ_p, has been fully ramped up. At $t \sim 46P_0$, two overdense blobs, associated with the edge mode, can be identified at the outer gap edge. At this time the gap has $\Delta\Sigma/\Sigma \simeq -0.3$, implying edge modes can develop *during* gap formation, since the steady state stable gap in $Q_o = 2$ reaches $\Delta\Sigma/\Sigma \simeq -0.6$. Edge modes are global and are associated with long trailing spiral waves with density perturbation amplitudes that are not small compared to that at the gap edge. This is clearly seen at $t \sim 50P_0$ when spiral shocks have already developed. We estimate that the edge modes have a growth rate consistent with predictions from linear theory and become non-linear within the time frame $46P_0 < t < 50P_0$.

The edge perturbations penetrate the outer gap edge and trail an angle similar to the planetary wake. Note the $m = 3$ disturbance near $r = 4$ in the snapshots taken at $t = 50P_0$ and $56P_0$. Two of the three overdensities, seen as a function of azimuth, correlate to the $m = 2$ mode in the outer disc, while the third overdensity adjacent to the planet correlates with the outer planetary wake. These features result from self-gravity and are unrelated to vortex modes.

The presence of both planetary wakes [with pattern speed $\simeq\Omega_k(5)$] and edge modes [with pattern speed $\simeq\Omega_k(5.5)$] of comparable amplitude enables direct interaction between them. Passage of edge mode spirals through the planetary wake may disrupt the former, explaining some of the apparently split-spirals in the plots. At $t \sim 64P_0$, Fig. 11 indicates that a spiral density wave feature coincides with the (outer) planetary wake. The enhanced outer planetary wake implies an increased negative torque exerted on the planet at this time. This effect occurs during the overlap of positive density perturbations. Assume that at a fixed radius the edge mode spiral has azimuthal thickness nh, where n is a dimensionless number. The time taken for this spiral to cross the planetary wake is $\Delta T = nh/|\Omega_k(5.5) - \Omega_k(5)| \simeq 0.06nP_0$. This is $\ll P_0$ for $n = O(1)$. We expect such crossing spiral waves to induce associated oscillations in the disc–planet torque.

The non-linear evolution of edge modes in disc–planet systems can give complicated surface density fields. The gap can become highly deformed. Fig. 11 shows that large voids develop to compensate for the overdensities in spiral arms, leading to azimuthal gaps ($t = 64P_0$). At $t = 72P_0$, the gap width ahead of the planet is narrowed by the spiral disturbance at the outer edge trying to connect to that at the inner edge. However, this gap-closing effect is opposed by the planetary torques which tend to open it.

A quasi-steady state is reached by $t \geq 76P_0$ showing an eccentric gap. For $t = 99.5P_0$ additional contour lines are plotted to indicate two local surface density maxima along the edge mode spiral adjacent to the planet. This spiral is disjointed around $r \sim 7.5$ where on traversing an arm, there is a minimum in $\Delta\Sigma/\Sigma$. Taking corotation of the spiral at $r_c = 5.5$ gives the outer Lindblad resonance of an $m = 2$ disturbance at $r = 7.2$, assuming a Keplerian disc. This is within the disjointed region of the spiral arm, which supports our interpretation that the disturbance at the edge is driving activity in the outer disc by perturbing it gravitationally and launching waves at outer Lindblad resonances.

7.3.1 Eccentric gaps

Several snapshots of the $Q_o = 1.5$ run taken during the second half of the simulation are shown in Fig. 12, which display the precession of an eccentric outer gap edge. Deformation of a circular gap into an eccentric one requires an $m = 1$ disturbance. By inspection of the form of the relative surface density perturbation for $Q_o = 1.5$ (see

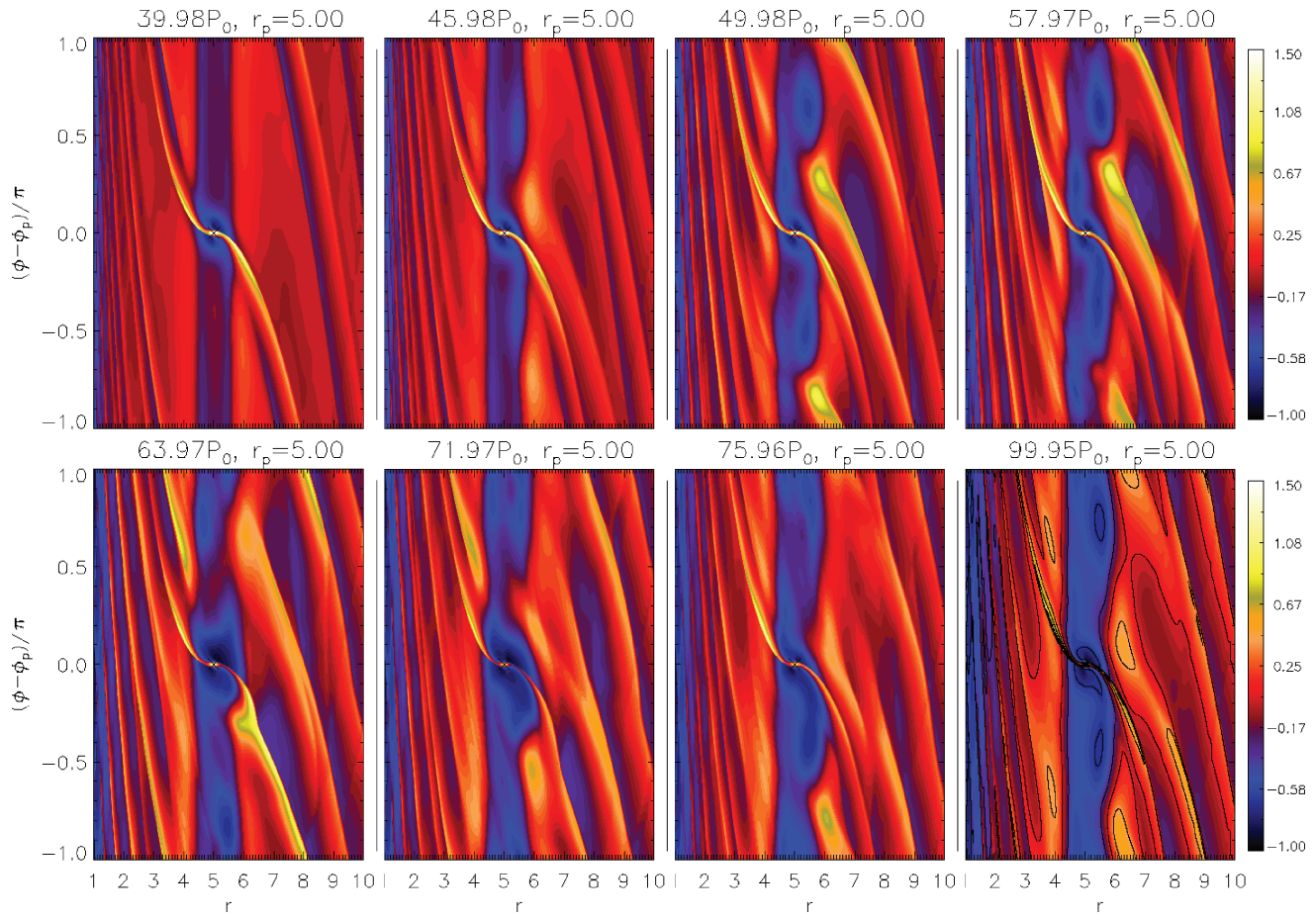

Figure 11. Evolution of the surface density perturbation (relative to $t = 0$) for the $Q_o = 1.5$ case from $t \sim 40P_0$ to $t \sim 100P_0$.

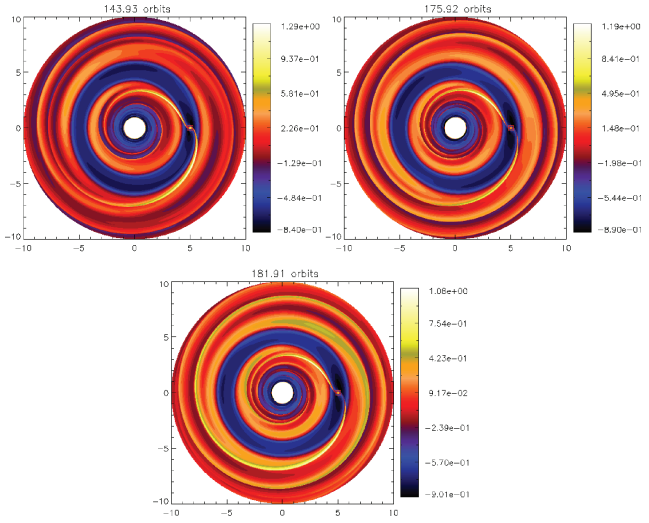

Figure 12. Precession of an eccentric gap in the presence of an $m = 2$ edge mode for a disc with $Q_o = 1.5$.

Fig. 10), we see that edge modes are associated with an eccentric outer gap edge. However, the inner gap edge remains fairly circular.

In their NSG disc–planet simulations, Kley & Dirksen (2006) found eccentric discs can result for planetary masses $M_p > 0.003M_*$ fixed on circular orbit. Our simulations show that edge modes in an SG disc can provide an additional perturbation to deform the gap into an eccentric shape for lower mass planets.

7.4 Evolution of the gap structure for $Q_o = 1.2$

With increasing disc mass, evolution of the gap profile away from its original form takes place. Fig. 13 illustrates the emergence of the $m = 3$ edge mode, expected from linear calculations, and subsequent evolution of the gap profile. By $t \sim 42P_0$ the $m = 3$ edge mode has already become non-linear, whereas for $Q_o = 1.5$ the $m = 2$ mode only begins to emerge at $t \sim 44P_0$, being consistent with linear calculations that the former has almost twice the growth rate of the latter. In Fig. 13 there is a $m = 3$ spiral at $r < r_p$ in phase with the mode for $r > r_p$. The gap narrows where they almost touch. These are part of a single mode. By $t = 44P_0$, shocks have formed and extend continuously across the gap. The self-gravity of the disc has overcome the planet's gravity which is responsible for gap opening. Edge mode spirals can be more prominent than planetary wakes.

The evolution of the form of the gap is shown in the one-dimensional averaged profiles plotted in Fig. 13. The quasi-steady profile before the onset of the edge mode instability is manifested at $t \sim 40P_0$. By $t = 44P_0$, the original bump at $r = 6$ has diminished. This bump originates from the planet expelling material as it opens a gap but is subsequently 'undone' by the edge mode as it grows and attempts to connect across the gap. This gap filling effect is possible as non-axisymmetric perturbation of the inner disc may be induced via the outer disc self-gravity. Note also in Fig. 13 at

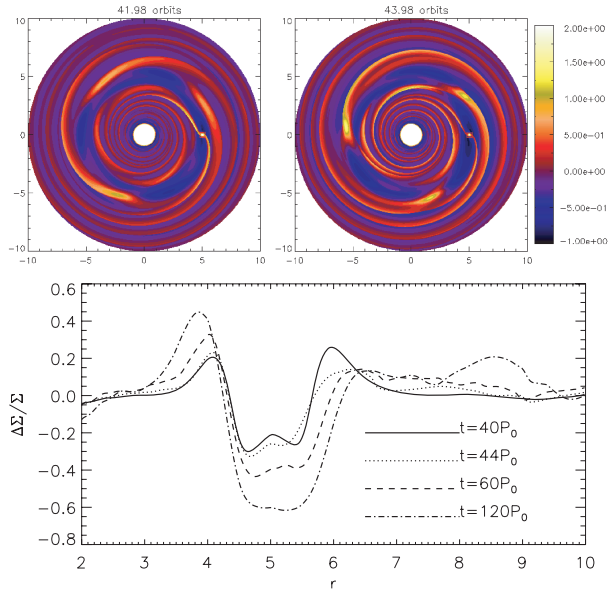

Figure 13. The disc model with $Q_o = 1.2$: emergence of the $m = 3$ edge mode (top) and evolution of the gap profile (bottom).

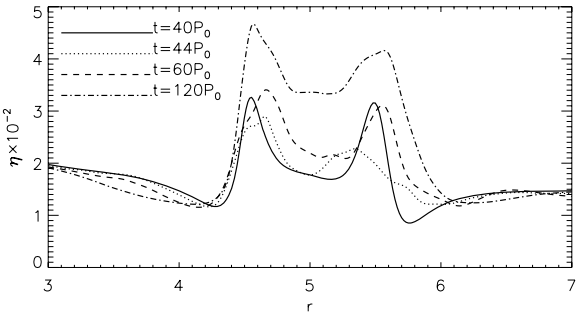

Figure 14. The disc model with $Q_o = 1.2$: evolution of the azimuthally averaged vortensity η profile in the region of the gap.

$t = 44P_0$ the increased surface density at $r \simeq 7.5$ implying radial redistribution of material. By $t = 120P_0$, the inner bump ($r = 4$) is similar to that produced by a standard gap-opening planet. By contrast, the outer bump cannot be maintained. The gap widens and there is an overall increase in surface density for $r > 6.5$. Note that there is no additional diminishing of the outer edge bump from $t = 44P_0 \rightarrow 120P_0$.

Fig. 14 illustrates the evolution of the vortensity profile within the gap. The development of edge modes temporarily reduces the amplitude of the vortensity maxima set up by the perturbation of the planet. This is particularly noticeable when the profiles at $t = 40P_0$ and $44P_0$ are compared. However, vortensity is generated through material passing through shocks induced by the gravitational perturbation of the planet (Lin & Papaloizou 2010). This provides a vortensity source that enables the amplitudes of vortensity maxima to increase again, as can be seen from the profile at $t = 60P_0$. The maxima remain roughly symmetric about the planet's location at $r_p \pm 2r_h$. Note a general increase in the co-orbital vortensity level caused by fluid elements repeatedly passing through shocks.

Although the $m = 3$ mode emerges first and is predicted to be most unstable from linear analysis, it does not persist. Linear theory can only predict the initially favoured mode. In the non-linear regime, Fourier analysis shows that $m = 2$ becomes dominant for $60P_0 \gtrsim t \gtrsim 145P_0$. We suspect this be due to finite viscosity acting over this time-scale. Larger m means shorter radial wavelengths, so over given a time-scale, diffusion and hence stabilization by viscosity is more effective than smaller m. The evolution is also complicated by the planetary wake which may gravitationally interact with edge modes. For $Q_o = 1.2$ we found that $m = 1$ becomes dominant at $t \gtrsim 145P_0$ when there appears to be a coupling between the outer planetary wake and the edge mode.

7.5 Mass transport

The similarity between the effects induced by an edge mode and an external perturber (planet) is further illustrated by the induced mass transport. Just as when a planet opens a gap, the development of an edge disturbance results in a radial redistribution of mass. We express the angular momentum transport in terms of an α viscosity parameter, defined through $\alpha = \Delta u_r \Delta u_\varphi / c_s^2$, where Δ here denotes deviation from the azimuthal average.

Angular momentum is also transmitted through gravitational torques. However, in practice we found this to be a small effect compared to transport due to Reynolds stresses. The discs here are not gravitoturbulent. The azimuthally averaged α parameters $\langle\alpha\rangle$ are shown in Fig. 15. The stable disc with $Q_o = 2$ has non-axisymmetric features entirely induced by the planetary perturbation and $\langle\alpha\rangle$ is localized about the planet's orbital radius giving a two-straw feature about r_p. This is typical of disc–planet interactions, and it provides a useful case for comparison with more massive discs to see the effect of edge modes and their spatial dependence.

Increasing self-gravity by adopting $Q_o = 1.5$, we see that angular momentum transport is enhanced around r_p and for $r < 4$, signifying the global nature of edge modes. Towards the inner boundary,

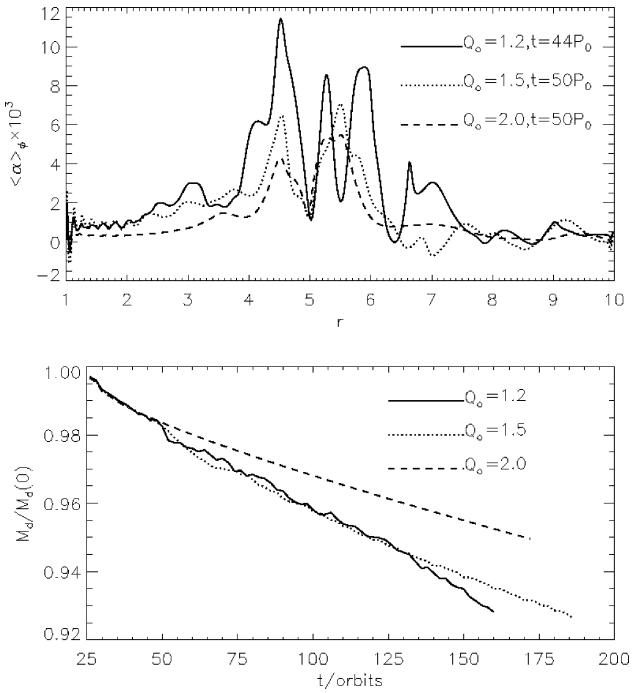

Figure 15. Top: azimuthally averaged Reynolds stresses associated with edge modes when they become non-linear ($Q_o = 1.2, 1.5$), the $Q_o = 2$ case is shown as a stable case with no edge modes for comparison. Bottom: evolution of the total disc mass for the three cases, scaled by initial mass.

$\alpha \simeq 10^{-3}$ being twice as large as for the case with $Q_o = 2$. The $Q_o = 1.5$ case also has enhanced wave-like behaviour for $r > 7$ as compared to $Q_o = 2.0$.

The $Q_o = 1.2$ case is more dramatic, with $\langle\alpha\rangle$ reaching a factor of 2–3 times larger than the imposed physical viscosity around the planet's location ($\nu/h^2 = 4 \times 10^{-3}$). This case displays a three-straw feature, with an additional trough at about $r = 5.5$, i.e. close to edge mode corotation, as if another planet were placed there. Although the $Q_0 = 1.5$ case also develops edge modes, there is no such additional trough at corotation. We suspect this may because the $Q_o = 1.2$ case is more unstable, developing an $m = 3$ mode (which produces three additional effective co-orbital planets placed at the gap edge), whereas the $Q_o = 1.5$ case develops a less prominent $m = 2$ mode. However, compared to $Q_o = 1.5$, $Q_o = 1.2$ has no enhanced transport away from the gap region.

Fig. 15 also shows the evolution of the disc mass as a function of time, $M_d(t)$. Viscous evolution leads to mass loss ($Q_o = 2.0$), and there is clearly enhanced mass loss if edge modes develop ($Q_o = 1.2, 1.5$). For $t \leq 50P_0$ the curves are identical, this interval includes gap formation and the emergence of edge modes. When edge modes become non-linear, mass loss is enhanced ($t \simeq 50P_0$), but there is no significant difference between $Q_o = 1.2$ and 1.5, which is consistent with the previous $\langle\alpha\rangle$ plots near the inner boundary. The $Q_o = 1.2$ curve is non-monotonic, though still decreasing overall, possibly because of boundary effects, and mass loss is enhanced once more around $t = 125P_0$. As Q_o is lowered, mass loss becomes less smooth, showing the dynamical nature of edge modes.

7.6 Additional effects

We briefly discuss effects of softening and viscosity on edge modes. To analyse mode amplitudes, the surface density field is Fourier transformed in azimuth giving $a_m(r)$ as the (complex) amplitude of the mth mode. We then integrate over the outer disc $r > 5$ to get $C_m = \int a_m \mathrm{d}r$ and consider $A = C_m/C_0$.

7.6.1 Softening

In linear theory, we found that edge modes cannot be constructed for gravitational softening parameters that are too large. We have simulated the $Q_o = 1.5$ disc with $\epsilon_{g0} = 0.6, 0.8, 1.0$.[2]

Fig. 16 shows the evolution of the running-time averaged $m = 2$ amplitudes. The evolution for all cases is very similar for $t \leq 40P_0$. Unstable modes develop thereafter, with increasing growth rates and saturation levels as ϵ_{g0} is lowered. $\epsilon_{g0} = 1$ displays either no growth or a much reduced growth rate. The gap for $Q_o = 1.5, \epsilon_{g0} = 1.0$ reaches a steady state similar to the stable case with $Q_o = 2.0, \epsilon_{g0} = 0.3$. This is consistent with the fact that as softening weakens, edge disturbances can no longer perturb the remainder of the disc via self-gravity, the residual non-axisymmetric structure being due to the planetary perturbation.

Interestingly, the decrease in saturation level in going from $\epsilon_{g0} = 0.3 \rightarrow 0.6$ is smaller than going from $\epsilon_{g0} = 0.6 \rightarrow 0.8$, which is in turn smaller than $\epsilon_{g0} = 0.8 \rightarrow 1.0$, despite a larger relative increase in softening in one pair than the next. This is suggestive of convergence for the non-linear saturation amplitude as $\epsilon_g \rightarrow 0$. This can be expected from convergence in linear theory (Section 6.4.1).

[2] It should be remarked that increasing ϵ_{g0} makes the Poisson kernel in simulations increasingly non-symmetric, so the earlier analytical discussion is less applicable.

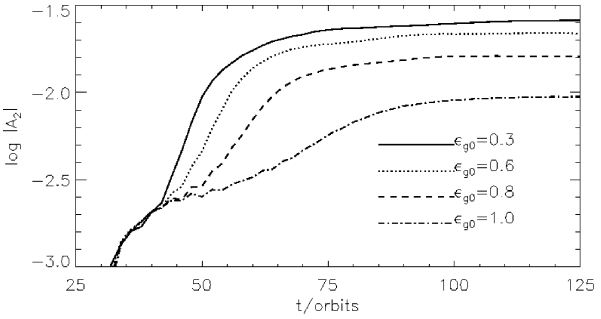

Figure 16. Running-time average of the surface density $m = 2$ Fourier amplitude, integrated over $r \in [5, 10]$ and scaled by the axisymmetric component, for the $Q_o = 1.5$ disc, as a function of the gravitational softening length.

However, at $\epsilon_{g0} = 0.3$ there are only five cells per softening length. To probe even smaller ϵ_{g0} requires prohibitively high-resolution simulations. Furthermore, since ϵ_g approximately accounts for the disc's non-zero vertical extent, very small softening lengths are not physically relevant to explore.

7.6.2 Viscosity

An important difference between edge-dominated modes and the already well-studied vortex modes in planetary gaps is the effect of viscosity on them. The standard viscosity $\nu = 10^{-5}$ prevents vortex mode development, but we have seen in Section 3 that such viscosity values still allow the $m = 2$ edge modes to develop. Vortices are localized disturbances and more easily smeared out than global spirals in a given time interval. Hence, vortex growth is inhibited more easily by viscosity than spirals.

We repeated the $Q_o = 1.5$ runs with a range of viscosities. Results are shown in Fig. 17. Generally, lowering viscosity increases the amplitudes of the non-axisymmetric modes. For $\nu \leq 10^{-6}$, the $m = 3$ mode emerges first (note that this is not in conflict with our linear calculations which used $\nu = 10^{-5}$ in its basic state), rather than $m = 2$. This is because lowering viscosity allows a sharper vortensity peak to develop in the background model, and thus has the same effect as increasing the disc mass. The latter can enable higher m edge modes.

Unlike vortex modes, even with twice the standard viscosity ($\nu = 2 \times 10^{-5}$), the edge mode develops and grows. It is only is suppressed when $\nu \geq 5 \times 10^{-5}$. However, this is because no vortensity maxima could be set up at the gap edges due to vortensity diffusion

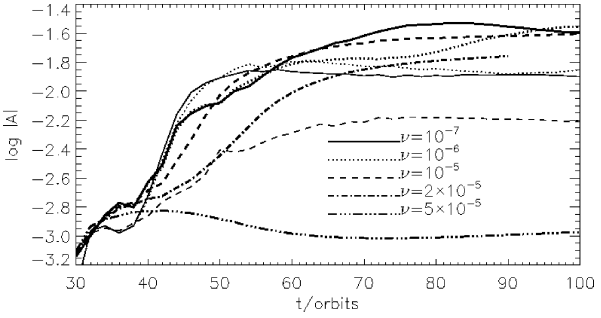

Figure 17. Non-axisymmetric disc modes for the $Q_o = 1.5$ disc as a function of applied viscosity ν. The Fourier amplitude has been integrated over $r \in [5, 10]$ and scaled by the axisymmetric amplitude and its running-time average plotted. Thick lines indicate the $m = 2$ mode and is plotted for all ν. Thin lines indicate the $m = 3$ mode and is plotted for $\nu = 10^{-7}$–10^{-5} only.

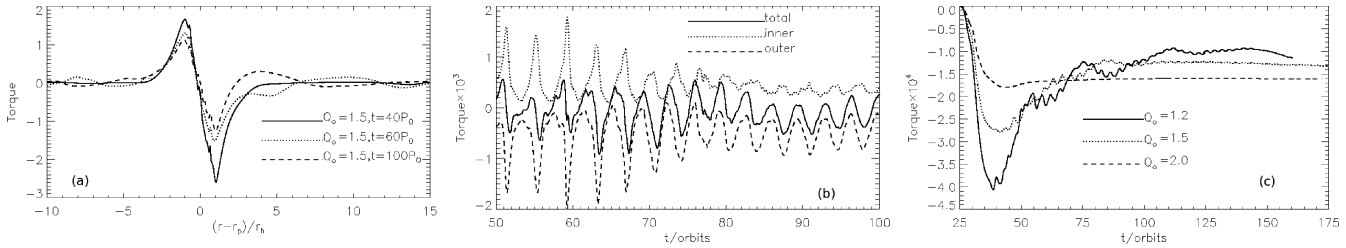

Figure 18. (a) Disc torque per unit radius acting on the planet. The length-scale $r_{\rm h} = (M_{\rm p}/3)^{1/3} r_{\rm p}$ is the Hill radius. (b) Time-dependent evolution of the disc–planet torques, the contribution from the inner disc (dotted curve), the outer disc (dashed curve) and the total torque (solid curve). (c) The running-time average of the total torque acting on the planet as a function of $Q_{\rm o}$.

in the co-orbital region making an almost uniform vortensity distribution in the gap. Since the necessary condition for edge modes is not achieved, no linear modes exist. This differs from the nature of the suppression of vortex modes, because in that case vortensity extrema that are stable can still be set up.

8 APPLICATIONS TO DISC–PLANET SYSTEMS

In order to focus on the issue of the stability of the dip/gap in the surface density profile induced by a giant planet, the planet was held on a fixed orbit. However, torques exerted by the non-axisymmetric disc on the planet will induce migration that could occur on a short time-scale (e.g. Pepliński et al. 2008). Such torques will be significantly affected by the presence of edge modes. Accordingly, we now discuss the torques exerted by a disc, in which edge modes are excited, on the planet.

8.1 Disc–planet torques

The presence of large-scale edge mode spirals of comparable amplitude as the planetary wake will significantly modify the torque on the planet, which in a stable disc is the origin of disc–planet torques. Here, we measure the disc–planet torque but still keep the planet on fixed orbit.

Fig. 18(a) shows the evolution of the torque per unit length for the $Q_{\rm o} = 1.5$ fiducial case. The torque profile at $t = 40P_0$, before the edge modes have developed significantly, shows the outer torque is larger in magnitude than inner torque, implying inward migration. This is a typical result for disc–planet interactions and serves as a reference. At $t = 60P_0$ and $100P_0$, edge modes mostly modify the torque contributions exterior to the planet, though some disturbances are seen in the interior disc ($r - r_{\rm p} < -5r_{\rm h}$, $r_{\rm h}$ being the Hill radius). The original outer torque at $+r_{\rm h}$, is reduced in magnitude as material redistributed to concentrate around the corotation radius $r_{\rm c}$ of the edge mode ($r_{\rm c}$ is $\sim 2r_{\rm h}$ away from the planet).

Edge modes can both enhance or reduce disc–planet torques associated with planetary wakes. Consider $r - r_{\rm p} \in [3, 5]r_{\rm h}$ in Fig. 18(a). Comparing the situation at $t = 60P_0$ to that at $t = 40P_0$, the torque contribution from this region is seen to be more negative. This is because an edge mode spiral overlaps the planetary wake, and therefore contributes an additional negative torque on the planet. However, at $t = 100P_0$, an edge mode spiral is just upstream of the planet, exerting a positive torque on the planet, hence the torque contribution from $r - r_{\rm p} \in [3, 5]r_{\rm h}$ becomes positive.

Differential rotation between edge modes and the planet produces oscillatory torques, shown in Fig. 18(b). Unlike typical disc–planet interactions, which produce inward migration, the total instantaneous torques in the presence of edge modes can be of either sign. For a disturbance with m fold symmetry the time interval between encounters with the planet is $\Delta T = (2\pi/m)/\delta\Omega$, where $\delta\Omega$ is the difference in angular velocity of the planet and the disturbance pattern. Approximating $\delta\Omega = |\Omega_{\rm k}(r_{\rm p}) - \Omega_{\rm k}(r_{\rm c})|$ with $r_{\rm c} = 5.46$ obtained from linear theory, we obtain $\Delta T = 4.04P_0$. Indeed, the oscillation period in Fig. 18 is $\simeq 4P_0$. Both inner and outer torques oscillate, but since the edge mode is more prominent in the outer disc, oscillations in the outer torque are larger, particularly after mode saturation ($t > 70P_0$).

We compare time-averaged torques in Fig. 18(c). In all three cases, on average a negative torque acts on the planet and its magnitude largest at $t = 40P_0$. Up to this point, the torque becomes more negative as the disc mass increases, as expected. However, development of edge modes makes the averaged torques more positive, and eventually reverses the trend with $Q_{\rm o}$. While the disc with $Q_{\rm o} = 1.5$ also attains steady torque values as in $Q_{\rm o} = 2$, the disc with $Q_{\rm o} = 1.2$ has significant oscillations even after time-averaging and remains non-steady at the end of the run.

Before the instability sets in, torques arise from wakes induced by the planet, and the outer (negative) torque is dominant. Edge modes concentrate material into spiral arms, leaving voids in between. Therefore, except when spiral arms cross the planet's azimuth, the surface density in the planetary wake is reduced because it resides in the void in between edge mode spirals. Hence the torque magnitude is reduced. If the reduction in surface density due to edge mode voids is greater than the increase in surface density scale produced by decreasing $Q_{\rm o}$, the presence of edge modes will, on average, make the torque acting on the planet more positive.

8.2 Outward scattering by spiral arms

If edge modes develop, planetary migration may be affected by them. Their association with gap edges inevitably affects co-orbital flows. While a detailed numerical study of migration is deferred to future work, we highlight an interesting effect found when edge modes are present. This is *outward* migration induced through scattering by spiral arms.

We restarted some of the simulations described above at $t = 50P_0$ allowing the planet to move in response to the gravitational forces due to the disc. The equation of motion of the planet is integrated using a fifth-order Runge–Kutta integrator.[3] Fig. 19 shows the orbital radius of the planet in discs with edge modes present. No obvious trend with $Q_{\rm o}$ is shown. This is because of oscillatory torques due to edge modes and the partial gap associated with a Saturn mass planet, which is associated with type III migration (Masset & Papaloizou 2003). The initial direction of migration can be inwards or

[3] Lin & Papaloizou (2010) found the same integrator to be adequate for studying migration induced through scattering by vortices.

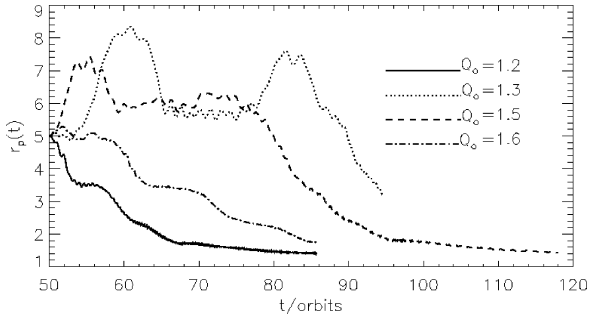

Figure 19. Orbital radius evolution in discs with edge-mode disturbances. The instantaneous orbital radius r_p is shown as a function of time.

outwards depending on the relative positioning of the edge mode spirals with respect to the planet at the time the planet was first allowed to move.

Outward migration is seen for $Q_o = 1.3$ and 1.5. For $Q_o = 1.5$, the planet migrates by $\Delta r = 2$, or 8.6 times its initial Hill radius, within only $4P_0$. This is essentially the result of a scattering event. Repeating this run with tapering applied to contributions to disc torques from within $0.6r_h$ of the planet, we found outward scattering still occurs, but limited to $\Delta r = 1$ and the planet remains at $r_p \simeq 6$ until $t = 128P_0$. Hence, while the subsequent inward migration seen for $Q_o = 1.5$ may be associated with conditions close to the planet, the initial outward scattering is due to an exterior edge-mode spiral.

For $Q_o = 1.3$ the planet is scattered to $r_p \simeq 8$. It is interesting to note that the subsequent inward migration for $Q_o = 1.5$ and 1.3 stalls at $r \simeq 6$, i.e. the original outer gap edge. The planet remains there for sufficient time for both gap and edge-mode development, and for $Q_o = 1.3$ a second episode of outward scattering occurs (it also occurs for $Q_o = 1.5$ to a small extent around $t = 70P_0$). Interaction with edge modes can affect the disc well beyond the original co-orbital region of the planet. In the case of outward migration, it may promote gravitational activity in the outer disc. However, boundary effects may be important after significant outward scattering.

The spiral arm–planet interaction for $Q_o = 1.5$ above is detailed in Fig. 20. A spiral arm approaches the planet from the upstream direction ($t = 50.9P_0$) and exerts positive torque on the planet, increasing r_p. The gap is asymmetric in the azimuthal direction with the surface density upstream of the planet being larger than that downstream of the planet ($t = 51.5P_0$). As the spiral arm passes through the planetary wake ($t = 51.5$–$52.1P_0$), the gap surface density just ahead of the planet builds up as an edge mode is set up across the gap. Note the fluid blob at $r = 5$, $\varphi - \varphi_p = 0.3\pi$. This signifies material executing horseshoe turns ahead of the planet. At the later time $t = 53.1P_0$, the planet has exited the original gap, in which the average surface density is now higher than before the scattering. Material that was originally outside the gap loses angular momentum and moves into the original gap. This material is not necessarily that composing the spiral.

We remark that migration through scattering by spiral arms differs from vortex scattering (Lin & Papaloizou 2010). Vortices form about vortensity minima, which lie further from the planet than vortensity maxima. We can assume vortensity maxima are stable in the case of vortex formation. This means the ring of vortensity maxima must be disrupted in order for vortices to flow across the planet's orbital radius for direct interaction. Edge modes themselves correspond to a disruption of vortensity maxima, hence direct interaction is less hindered than in the vortex case. Furthermore, a vortex is a material volume of fluid. Vortex-planet scattering results in an orbital radius

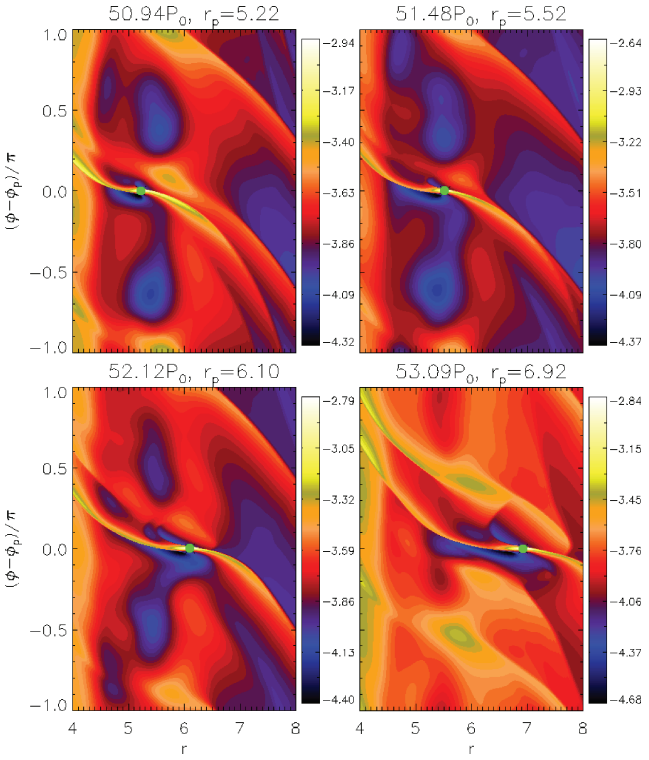

Figure 20. Interaction between an edge-mode spiral and the planet (green dot), causing the latter to be scattered outwards. log Σ is shown.

change of both objects. A spiral pattern is generally not a material volume of fluid. It can be seen in Fig. 20 that the local max(Σ) in the spiral remains at approximately the same radius before and after the interaction. The spiral as a whole does not move inwards. In this case, the spiral first increases the planet's orbital radius, thereby encouraging it to interact with the outer gap edge and scatter that material inwards.

9 SUMMARY AND DISCUSSION

We have studied instabilities associated with a surface density gap opened by a giant planet embedded in a massive protoplanetary disc. Vortex producing instabilities, associated with local vortensity minima, have previously been found to occur in weakly or NSG discs. However, edge modes that are associated with local vortensity maxima and require sufficiently strong self-gravity were found in the more massive discs that we have focused on in this paper. These have very different properties in that they are global rather than local and are associated with large-scale spiral arms as well as being less affected by viscous diffusion.

For our disc–planet models with fixed planet mass $M_p = 3 \times 10^{-4} M_*$ and fixed disc aspect ratio $h = 0.05$, we found edge modes to develop for gap profiles with an average Toomre $Q \lesssim 2$ exterior to the planet's orbital radius. In the unperturbed smooth disc, this corresponds to $Q \lesssim 2.6$ at the planet's location and $Q \lesssim 1.5$ at the outer boundary, or equivalently a disc mass $M_d \gtrsim 0.06 M_*$. For fixed M_d, non-linear simulations show edge modes develop readily for uniform kinematic viscosity $\nu \lesssim 2 \times 10^{-5}$ and self-gravitational softening $\epsilon_g \lesssim 0.8H$.

A theoretical description of edge modes was developed. The edge mode is interpreted as a disturbance associated with an interior vortensity maximum that requires self-gravity to be sustained,

and which further perturbs the remainder of the disc through its self-gravity causing the excitation of density waves. Linear calculations confirm this picture and also show the expected strengthening of the instability when self-gravity is increased via the disc mass and the expected weakening of the instability through increasing gravitational softening.

Hydrodynamic simulations of disc–planet interactions were performed. We confirm earlier suggestions by Meschiari & Laughlin (2008), who found gravitational instabilities associated with their prescribed gap profiles without a planet. Indeed, we found that edge modes can develop for gaps self-consistently opened by introducing a planet into the disc. However, our models showed their development *during* gap formation of a Saturn mass planet, when the gap only consists of a 20–30 per cent deficit in surface density relative to the unperturbed disc. This is much less shallow than the Jovian mass planetary gaps considered in Meschiari & Laughlin (2008), which are typically associated with a 90 per cent deficit. We found edge modes to exist for disc models extending to a distance twice as large as the planets orbit with masses $M_{\rm d} \gtrsim 0.06 M_*$, again this is less massive than required in Meschiari & Laughlin (2008), but more massive than those typically used in modelling protoplanetary discs.

We remark that in the disc model of Meschiari & Laughlin, the local Q maximum occurs at the gap centre, presumably where the planet would lie. However, gap opening by the planet may disrupt the disturbance associated with this extremum, which is required to induce perturbations in the smooth disc. It is then unclear if the edge mode could be set up. Furthermore, since corotation lies at their gap centre, there would be no relative motion between spirals and the planet of the kind seen in our simulations. The case of Jovian mass planets will be explored in a future work. We note in that case, significant disruption of the disc, leading to fragmentation, may occur (Armitage & Hansen 1999; Lufkin et al. 2004).

Our simulations show edge modes with $m = 2$ and 3. They have surface density maxima localized near the outer gap edge, but the disturbance they produce extends throughout the entire disc, consistent with analytical expectation and linear calculations. One important difference between edge modes and the vortex forming modes in NSG discs is that the latter require low viscosity. The typical viscosity values adopted for protoplanetary discs, $\nu = 10^{-5}$, will suppress vortex formation but not edge modes associated with self-gravity.

We have also considered the effect of edge modes on planetary migration. They produce large oscillations in the disc torques acting on the planet, which can be positive or negative. The presence of edge modes reverses the trend of the time-averaged disc-on-planet torque as a function of disc mass: the torque being more positive as disc mass increases. Direct interaction between the planet and a spiral arm associated with an edge mode is possible. These should be more prominent in the disc section with the lower Q. In practise this corresponds to $r > r_{\rm p}$, so the planet tends to interact with the outer spirals, which results in a scattering of the planet outwards.

9.1 Outstanding issues

This work has been motivated by the wish to obtain a better understanding of disc–planet interactions in massive discs. In particular, a proper discussion of phenomena such as type III migration requires incorporation of self-gravity which we have undertaken here. We have studied model discs which are unstable through the development of edge modes when a dip/gap is produced by an embedded giant planet but which would be stable without the planet.

Our first calculations show that the disc torques acting on the planet in the presence of edge modes change sign and are highly time-dependent. Thus migration is unpredictable. A statistical approach may be ultimately required to assess the likelihood of reduced inward migration in practice. Furthermore, development of edge modes during gap formation indicates their importance for planets before they reach a Jovian mass. This leads to the possibility of type III migration, which is self-sustained and can be in either direction.

Issues such as what is the appropriate softening length to use in a two-dimensional simulation and the treatment of material inside the Hill radius will be important for more thorough understanding of planetary migration in the presence of large-scale spiral arms in massive discs. Resolving these requires improved numerical modelling which will be presented in a future work.

ACKNOWLEDGMENTS

MKL acknowledges support from St John's College, Cambridge, the Isaac Newton Trust and an Overseas Research Award.

REFERENCES

Adams F. C., Ruden S. P., Shu F. H., 1989, ApJ, 347, 959
Armitage P. J., Hansen B. M. S., 1999, Nat, 402, 633
Baruteau C., Masset F., 2008, ApJ, 678, 483
Binney J., Tremaine S., 1987, Galactic Dynamics. Princeton Univ. Press, Princeton, NJ
de Val-Borro M. e., 2006, MNRAS, 370, 529
de Val-Borro M., Artymowicz P., D'Angelo G., Peplinski A., 2007, A&A, 471, 1043
Godon P., 1996, MNRAS, 282, 1107
Goldreich P., Tremaine S., 1979, ApJ, 233, 857
Kley W., Dirksen G., 2006, A&A, 447, 369
Koller J., Li H., Lin D. N. C., 2003, ApJ, 596, L91
Li H., Finn J. M., Lovelace R. V. E., Colgate S. A., 2000, ApJ, 533, 1023
Li H., Colgate S. A., Wendroff B., Liska R., 2001, ApJ, 551, 874
Li H., Li S., Koller J., Wendroff B. B., Liska R., Orban C. M., Liang E. P. T., Lin D. N. C., 2005, ApJ, 624, 1003
Li H., Lubow S. H., Li S., Lin D. N. C., 2009, ApJ, 690, L52
Lin D. N. C., Papaloizou J. C. B., Kley W., 1993, ApJ, 416, 689
Lin M., Papaloizou J. C. B., 2010, MNRAS, 405, 1473
Lin M., Papaloizou J. C. B., 2011, MNRAS, in press (doi:10.1111/j.1365-2966.2011.18798.x) (this issue)
Lovelace R. V. E., Li H., Colgate S. A., Nelson A. F., 1999, ApJ, 513, 805
Lufkin G., Quinn T., Wadsley J., Stadel J., Governato F., 2004, MNRAS, 347, 421
Lynden Bell D., Ostriker J. P., 1967, MNRAS, 136, 293
Masset F., 2000, A&AS, 141, 165
Masset F. S., 2002, A&A, 387, 605
Masset F. S., Papaloizou J. C. B., 2003, ApJ, 588, 494
Mayor M., Queloz D., 1995, Nat, 378, 355
Meschiari S., Laughlin G., 2008, ApJ, 679, L135
Ou S., Ji J., Liu L., Peng X., 2007, ApJ, 667, 1220
Paardekooper S., Lesur G., Papaloizou J. C. B., 2010, ApJ, 725, 146
Papaloizou J., Lin D. N. C., 1984, ApJ, 285, 818
Papaloizou J. C. B., Lin D. N. C., 1989, ApJ, 344, 645
Papaloizou J. C. B., Pringle J. E., 1985, MNRAS, 213, 799
Papaloizou J. C., Savonije G. J., 1991, MNRAS, 248, 353
Pepliński A., Artymowicz P., Mellema G., 2008, MNRAS, 386, 179
Sellwood J. A., Kahn F. D., 1991, MNRAS, 250, 278
Stone J. M., Norman M. L., 1992, ApJS, 80, 753
Yu C., Li H., Li S., Lubow S. H., Lin D. N. C., 2010, ApJ, 712, 198
Zhang H., Yuan C., Lin D. N. C., Yen D. C. C., 2008, ApJ, 676, 639

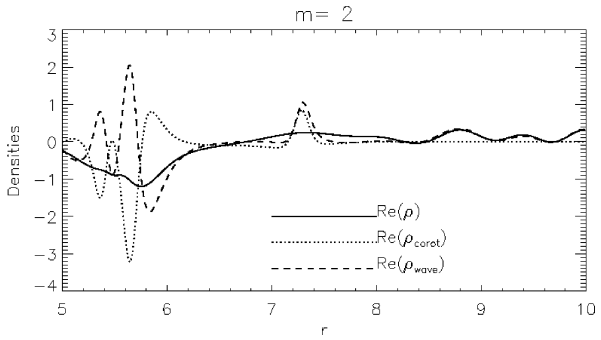

Figure A1. The thermal and gravitational energy (TGE) per unit length (solid) computed from eigensolutions for the fiducial case $Q_o = 1.5$ from linear theory. The contributions from the vortensity term (dotted line) and other terms (dashed line), defined by equation (67), are also shown. The relatively small bumps near $r = 7$ are numerical. However, as they are the same magnitude for both contributions, incorporating them does not affect conclusions concerning the overall energy balance in the system.

APPENDIX A: ENERGY DENSITIES IN LINEAR THEORY

In Section 6 we defined and discussed the TGE and the different contributions to it. There, a factor D^2 was applied to the TGE per unit length and to each contributing term in order to overcome numerical difficulties associated with Lindblad resonances. For completeness, we provide here a discussion of the behaviour of the TGE and the various contributions to it considered without the additional factor D^2.

Fig. A1 shows very similar features to the corresponding curves discussed in Section 6.2. Re(ϱ) is negative around corotation, again signifying a self-gravity-driven edge mode. Beyond $r \simeq 6.4$, Re(ρ) becomes positive due to pressure. The most extreme peak at $r \simeq 5.6$ coincides with the background vortensity edge (Fig. 4). The negative contribution from the corotation term is larger in magnitude than that from the positive wave term, resulting in $\mathrm{Re}(\rho) < 0$. The corotation radius $r_c < 5.6$, making the shifted frequency $\bar{\sigma}_R(5.6) = m[\Omega(5.6) - \Omega(r_c)] < 0$. Also at $r = 5.6$, $\mathrm{d}(\eta^{-1})/\mathrm{d}r > 0$, resulting in $\mathrm{Re}(\rho_{\mathrm{corot}}) < 0$ at this point.

Re(ρ_{wave}) and Re(ρ_{corot}) have relatively small spurious bumps around $r = 7.2$ that are associated with the outer Lindblad resonance and are numerical. The eigenfunctions W, Φ' are well defined without singularities there, but evaluation of ρ_{corot} and ρ_{wave} involves division by D, which can amplify numerical errors at Lindblad resonances where $D \to 0$. Integrating Re(ρ) over [5, 10], we find $U < 0$, the TGE is negative, which means gravitational energy dominates over the pressure contribution for $r \geq 5$. Integrating the contributions to the TGE separately, we find

$$U_{\mathrm{corot}} \equiv \mathrm{Re}\int_5^{10} \rho_{\mathrm{corot}}\mathrm{d}r \simeq -0.94|U|,$$

$$U_{\mathrm{wave}} \equiv \mathrm{Re}\int_5^{10} \rho_{\mathrm{wave}}\mathrm{d}r \simeq -0.077|U|.$$

The integration range includes the OLR and thus the spurious bumps in ρ_{wave} and ρ_{corot}. However, we still find that U approximately equals $U_{\mathrm{corot}} + U_{\mathrm{wave}}$. The correct energy balance is still maintained despite being subject to possible numerical error due to the diverging factor $1/D$. The above means that, aside from the spurious bumps, the energy density values for $r \in [5, 10]$ may still be used to interpret energy balance. We find $|U_{\mathrm{corot}}/U| \sim 0.9$, so as before the TGE is predominantly accounted for by the vortensity term.

This paper has been typeset from a TEX/LATEX file prepared by the author.